الجينوميكس والمعلوماتية الحيوية

Genomics & Bioinformatics

رئيس التحرير

أ.د. أحمد يوسف المتينى

كلية الزراعة – جامعة الإسكندرية

المشاركون فى التحرير

أ.د. سناء أحمد رياض
كلية الزراعة
جامعة الإسكندرية

د. آمال أحمد عبد العزيز
معهد الهندسة الوراثية
مدينة السادات جامعة المنوفية

د. ياسر محمد مبروك
كلية الزراعة .جامعة الإسكندرية

د. أحمد محمد الشهاوى
كلية الزراعة .جامعة الإسكندرية

م. أمير السعيد يسن

كلية الزراعة – جامعة الإسكندرية

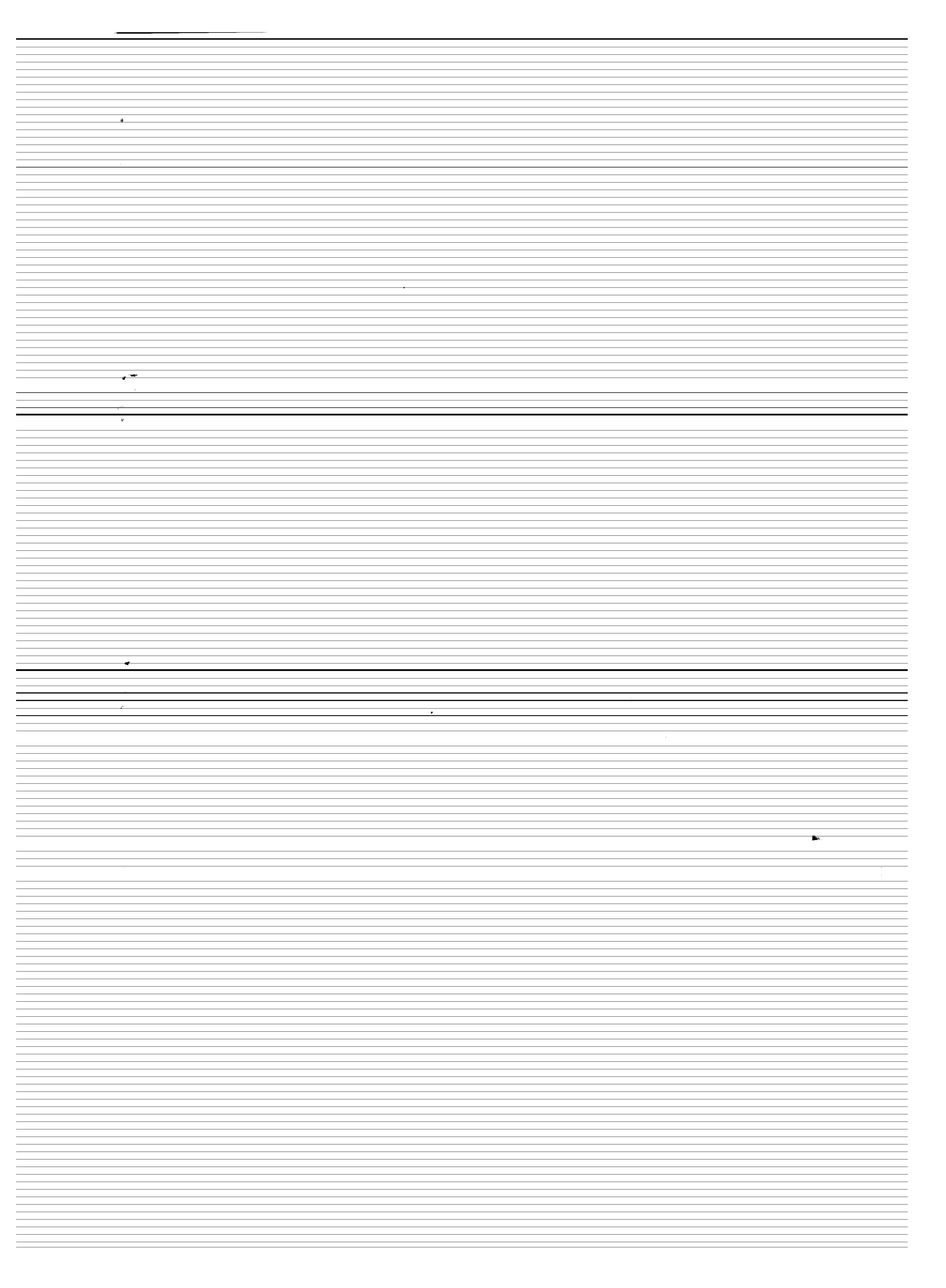
2006

مكتبة بستان المعرفة

طباعة ونشر وتوزيع الكتب

كفر الدوار ـ الحدائق بجوار نقابة التطبيقيين

☎ : ٠٤٥/٢٢٢٤٢٢٨ & ١٢١١٥١٢٣٧



الجينوميكس والمعلوماتية الحيوية

Genomics & Bioinformatics

اسم الكتاب	الجينوميكس و المعلوماتية الحيوية
اسم المؤلف	أ.د. أحمد يوسف المتيني
رقم الإيداع	٢١٩٠٢ /٢٠٠٥
الترقيم الدولي	I.S.B.N 977-393-057 -2
الطبعة	الأولى
الناشر	مكتبة بستان المعرفة
	كفر الدوار الحدائق ٦٧ ش الحدائق بجوار نقابة التطبيقيين
	☎: ٠٤٥/٢٢٢٤٢٢٨ الإسكندرية ٠١٢١١٥١٢٣٧
	Email: bostan_elma3rafa @ yahoo.com

Preface تمهيد

يعتبر الكثير من المفكرين أن اكتشاف التركيب الجزيئى للمادة الوراثية (الا وهو الـ DNA) من أعظم إنجازات القرن العشرين قاطبة، فبه وما تلاه من دراسات اصبحنا نعرف لغة الحياة. فالنيوكلوتيدات الأربع هى كل حروف تلك اللغة، تتابع فى ثلاثيات كل منها يشفر لحمض أمينى بعينه، ومن تلك الأحماض تتكون توليفات لانهائية من البروتينات عصب حياة الكائن. فإذا ما أعتبر المفكرين ذلك بأنه إنجازا إنسانيا، فهو لنا نحن الوراثيون ثورة قلبت علوم الحياة ودفعتها إلى آفاق لم نحلم بها من قبل ومع ظهور تقنيات الهندسة الوراثية أصبحنا قادرين على تطويع الوراثة تبعا لرغباتنا، واصبح من الصعب على المتخصصين أن يلاحقوا أو يتابعوا مكتشفات علوم الوراثة الجزيئية خصوصا والبيولوجيا الجزيئية عموما، وأصبح البحاث فى تلك المجالات كل يعمل بعيدا عن الآخر مثل الجزر المتباعدة فى المحيط، مما أفقدنا القدرة على استخلاص الكليات وغرقنا فى بحر التفصيلات الصغيرة غير المترابطة، وظل الحال على ما هو عليه إلى أن قامت مجموعة من علماء الأحياء والرياضة فى مطلع عام ١٩٧٩ فى أجتماع جمعهم فى أحد قاعات جامعة روكفلر Rockefeller University بمدينة نيويورك حيث دعوا إلى إنشاء قاعدة معلومات لحفظ تتابعات الـ DNA، ومن بعدها مباشرة وفى السنوات القليلة التالية ظهر اتجاه جديد لمحاولة الربط ما بين علوم الحياة وعلوم الكمبيوتر والبرمجة ويعرف هذا المجال باسم المعلوماتية الحياتية Bioinformatics، هذا المجال الجديد أدخل تغيرات جذرية فى المناهج العلمية لدراسة علوم الحياة خصوصا البيولوجيا الجزيئية، وقد تسابقت الجامعات والمعاهد العلمية فى دول العالم المتقدم إلى إضافة هذا المجال الجديد إلى مناهجها حتى تواكب التغير المأمول. ونحن هنا لسنا فى صدد دراسة الـ Bioinformatics من حيث اسسها النظرية، فذلك يندرج تحت علوم الكمبيوتر والهندسة، بل يهمنا أن نعلم هنا التطبيقات والتقنيات التى يمكن أن تفيدنا فى دراستنا للوراثة الحديثة. ويمكننا أن نسوق هنا، مثالا طريفا يؤكد أهمية المعلوماتية الحياتية فى مجال العلوم البيولوجية الحديثة. فقد اختير فى مطلع عام ٢٠٠٠ طالب الدراسات العليا James Kent الذى يدرس فى معمل العالم البيولوجى الشهير Zahler بجامعة كاليفورنيا ليكون بطلا قوميا بواسطة جريدة New York Time لنجاحه فى عمل برنامج للكمبيوتر أطلق عليه اسم

"GigAssembler". هذا البرنامج مكن الفريق القومى من مشروع الجينوم البشرى من حل تتابعات حوالى ٤٠٠,٠٠٠ قطعة متداخلة من الـ DNA (contigs) كانت مازالت باقية لم تحل بعد، هذا البرنامج أتمها فى غضون أربعة أسابيع فقط. هذا الإنجاز مكن الفريق القومى من ملاحقة فريق القطاع الخاص بالمشروع والممثل بشركة Celera Genomics، و التى كانت قد هددت بالاحتفاظ بالنتائج لنفسها لاستغلالها تجاريا. وبذلك تكللت الجهود المشتركة فى الإعلان عن نتائج المشروع الأولية بواسطة كلا الجانبين فى٢٦ يوليو سنة ٢٠٠٠.

وبعيدا عن تعقيدات المعادلات الرياضية وبرامج الكمبيوتر فالمعلوماتية تبدأ أساسا من وجود قواعد للمعلومات أو بمعنى أخر من وجود نظام لتبويب وتنظيم المعلومات، ومن خلال هذا المفهوم فيمكن أعتبار المكتبات (الكتب والدوريات والمجلات والصحف)،هى المكان حيث تجمع المطبوعات وتبوب وتنظم حسب لمعايير عديدة متباينة. بهذا المفهوم المعلوماتى، كان لمدينتنا (مدينة الإسكندرية Alexandria) شرف السبق فمنذ حوالى ٢٠٠٠ سنة مضت كانت الأسكندرية منارة للعلم ومركز ثقافى وتجارى متميز بجنوب شرق البحر الأبيض المتوسط، وكانت مكتبتها الشهيرة هى الأولى من حيث تنظيم وتبويب المعلومات والمخطوطات التى جمعت من كافة الحضارات مثل اليونانية و الأشورية و الفارسية والمصرية والعبرية والهندية وغيرها. فإذا ما كان لنا شرف السبق فوجب علينا الآن اللحاق بالدرب، فياليتنا نفيق ونلحق ما فاتنا.

وكان من النتائج المباشرة لتغلغل المعلوماتية فى العلوم البيولوجية أن ظهر فى السنوات العشر الماضية مجموعة من العلوم الجديدة والتى تتميز جميعها بالمقطع (omics-) أو ما يعرف مجازا بعلوم "الأومكس" مثل الجينومكس (Genomics) والبروتيومكس (Proteimics) و الفينومكس (Phenomics) وغيرها. ويمكن تعريف الجينومكس بأنه الفرع من العلم الذى يهتم بدراسة المحتوى الوراثى ككل لكائن ما.

ومصطلح جينوم genome معروف فى الوراثة منذ أوائل الثلاثينات من القرن الماضى بين رجال الوراثة السيتولوجية cytogenetics، ويعنى العدد الأحادى من الكروموسومات haploid. وقد نوه العالم C. P. Blacker فى كتابة المعنون *Eugenics*

المنشور سنة ١٩٥٢ بأن ظهور مصطلحات وراثية مثل المجاميع الجينية والجينوم التى غيرت من المفاهيم الوراثية السابقة من حيث التركيز على الجين بصفته وحدة التوارث إلى الأهتمام بالوحدات الأكبر مثل الجينوم. ويمكن أيضا تعريف الجينومكس بأنه تقنيات "السلسلة" sequencing جزيئات الـ DNA الكلية (الجينومية) لكائن ما، ولكن أمكانية تطبيق تلك التقنيات لم تكن ممكنة قبل مجهودات عالم الوراثة العظيم Sanger سنة ١٩٧٧ حيث تمكن من سلسلة الجينوم الكامل للفيروس البكتيرى الصغير *phiX174* والبالغ قدره ٥٣٨٦ زوج من النيكلوتيدات فقط، وقد حصل على جائزة نوبل فيما بعد على إنجازه هذا. ومع أن الطرق التى اتبعها كانت بطيئة (بمفهومنا الحالى) إلا أنها كانت البداية التى فتحت الباب على مصرعية نحو بذوغ فجر الجينومكس، خصوصا بعد الأعلان عن استكمال سلسلة الجينوم الكامل للبكتيريا *Haemophilus influenzae* (١,٨٣٠,٠٠٠ زوج من النيوكلوتيدات) سنة ١٩٩٥.

كذلك وجب التنويه، بأننا قد استعملنا المصطلحات الخاصة بهذا الموضوع دون محاولة تعريبها، بل هى المصطلحات الأجنبية مكتوبة بحروف عربية، ومع قناعتى بعدم صحة هذا الاتجاه إلا أن واقع الأمر فرضه علينا. فهذا المجال حديث للغاية، بل انه يعتمد أساسا على شبكة الأنترنت فى الدراسة والتحصيل فكان لزاما أن يعرف الدارس المصطلح الذى يمكنه من الأتصال بتلك المصادر العالمية بسهولة. تاركين الباب فى المستقبل، بعد أن ترسخ تلك المفاهيم الجديدة فى ذهن القارئ العربى، لكي نعربها دون الإخلال بالمعنى ولكن تلك مهمة لاحقة قد تجد من يتصدى لها. ومع ذلك فقد حاولنا تعريب عدد من المصطلحات الوراثية والتطورية الجديدة فيما هو متاح، ولكن نستميح القارئ عذرا فى إحتمال توارد اكثر من تعريب لنفس المصطلح لتعدد المشاركين فى إنجاز هذا العمل.

ولا يفوتنا تقديم جليل الشكر للدكتور محمد عبد الفتاح ياقوت لعظيم مساعداته لإخراج وطباعة هذا العمل فى أكمل وجه ممكن.

وختاما أتمنى أن يلقى هذا الجهد المتواضع اهتمام الدارسين وأن يكون سبيلهم للإضافة والبحث. والله من وراء القصد،،،

أحمد يوسف المتينى

الإسكندرية فى ديسمبر ٢٠٠٥.



المحتويات

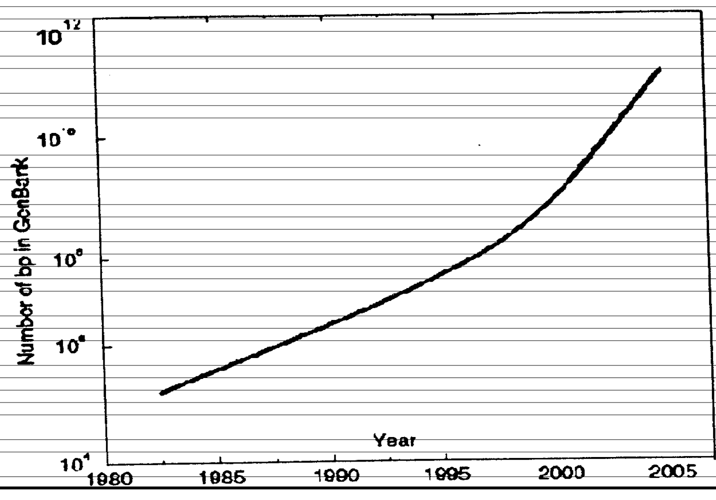
Contents

١. المعلومات الوراثـــية
Genetic Information
إعداد: أحمد المتيــنى

ركز علماء الوراثة، منذ أربعينيات القرن الماضى وحتى الآن، جل جهدهم وفكرهم فى التعرف على الطبيعة الجزيئية للجينات molecular nature of genes، حيث قتل الموضوع بحثا من كافة جوانبه تقريبا وتراكم لدينا كم هائل من النتائج أصبح الإلمام بكافة جوانبها من شبه المستحيل حتى على المتخصصين. وزاد من تعقيد الوضع القائم التقدم الهائل خلال السنوات العشرة الأخيرة فى الأنتهاء من دراسة جينومات العديد من الكائنات ومن ضمنها الإنسان، والجدول التالى يحدد عدد الجينومات التى تم معرفتها حتى سنة ٢٠٠٢.

السنة	عدد الجينومات المدروسة لكل سنة	العدد التراكمي
١٩٩٤	-	-
١٩٩٥	٢	٢
١٩٩٦	٢	٤
١٩٩٧	٥	٩
١٩٩٨	٨	١٧
١٩٩٩	١٣	٣٠
٢٠٠٠	٢٣	٥٣
٢٠٠١	٤٢	٩٥
٢٠٠٢	١٠٠	١٩٥

وقد يبلغ عدد الكائنات التى تم التعرف على جينوماتها حتى نهاية عام ٢٠٠٥ ضعف هذا الرقم مما زاد بشكل مهول حجم البيانات المتاحة. شكل (١-١) يبين معدل الزيادة فى حجم البيانات المخزنة بالبنوك الجينية gene banks من عام ١٩٨٠ حتى نهاية عام ٢٠٠٥، مما يوضح بصورة جلية قدر المشكلة التى يعانى منها البيولوجيون عموما والوراثيون خصوصا. هذه الصعوبة دفعت العديد من العلماء والباحثين فى التفكير فى مخرج من هذه المشكلة. وكان الحل المأمول هو الأستعانة بعلوم الكمبيوتر والبرمجة.



شكل (١-١) : عدد أزواج النيوكلوتيدات المخزنة فى البنوك الجينية خلال الخمسة وعشرون سنة الماضية.

الوراثة مثلها مثل أى طراز آخر من المعلومات information لابد من أن تعتمد على لغة language أو وسيلة ما للتواصل ونقل المعلومة، لذلك فالمعلومات الوراثية لها لغتها الخاصة (ثلاثيات الشفرة الوراثية -- genetic code) والتى لاتختلف كثيرا عن اللغات الحية (مثل الأنجليزية والفرنسية والعربية.... وغيرها) ولاتختلف أيضا عن لغة الكمبيوتر machine language كثيرا. والمعلومات الوراثية لا تختلف عن أى لغة أخرى، فهى تتميز بالصفات المعلوماتية العامة مثل:

(١) أنها معلومات رقمية (digital signals) ؛ النيوكلوتيدات الأربع تكون ثلاثيات الشفرة المستقلة والتى ليس بينها تداخل (triplet-nucleotides - codons) (non-overlapping) حيث تمثل الحروف الهجائية لبناء البروتين

(٢) المعلومات تمثل مصفوفات خطية (أوتار) (linear strings) مكونة غالبا من وحداتها الكودون المصفوفة والمتجاورة

(٢) تمتاز أيضا بوجود نظام للتحقق والتصفية (filtering) للتعرف على المعلومات ذات المدلول (sense-information) أو عديمة المدلول أو الشوشرة (non-sense or noise).

١- طبيعة الجينات Nature of Genes.

الوراثة هى العلم الذى يختص بدراسة إنتقال الصفات (المظهرية) من أفراد النوع الواحد إلى أبنائهم من خلال التكاثر الجنسى. ومنذ مندل وحتى الآن فإن مفهومنا عن طبيعة العامل الوراثى (الجين) المسئول عن نقل الصفات من الآباء إلى الأبناء، مبهم وغير محدد ! ففى القرنين السابع عشر والثامن عشر، ساد الأعتقاد بنظرية "سبق التكوين" preformation theory حيث كانوا يعتقدون بتكون الفرد كاملا قبل المولد (أى عدم الأعتراف بوجود الجين)، ولكن مع مطلع القرن التاسع عشر نادى عدد من العلماء وعلى رأسهم دارون و ويزمان بوجود العامل الوراثى كوحدة مادية (particulate theory) ولكن أعتقدوا بأن هذه العوامل يمكن المزج بينها، هنا جاء الدور العظيم لمندل حيث بين أن هذه العوامل (الجينات) لاتمتزج ولكن تربطها علاقات من السيادة والتنحى. وبإعادة أكتشاف تجارب مندل مع مطلع القرن العشرين، ساد المفهوم المندلى للجين أو ما يسمى بالجين الكلاسيكى (classical gene)، ولكن يجب ان نؤكد هنا أن المفهوم الكلاسيكى للوراثة كان جل أهتمامه التنبؤ (prediction) بالأشكال المظهرية المتوقعة بفرضية وجود الجينات المسئولة عنها دون معرفة طبيعتها (وهذا كان أفتراضا نظريا بحتا)، أى الجين هى بمثابة "صندوق أسود" حيث تمثل عوامل الأباء المدخلات inputs بينما الشكل الظاهرى يمثل المخرجات outputs، بينما محتوى هذا الصندوق فمجهول تماما، مع العلم أن الرابطة بين الشكل الظاهرى والجينات لابد أن تتم من خلال عمليات التشكل والنمو (development) تبعا للفكر المندلى ذاته. هذا التجاهل لطبيعة الجين المادية تداعا مع منتصف القرن العشرين مع التوجه لدراسة البيولوجيا الجزيئية عموما والوراثة الجزيئية خصوصا بفضل مجهودات عدد من علماء الوراثة العظام الذين ساهموا بدون شك فى تغيير الكثير من المفاهيم الوراثية التى سادت لأكثر من قرن من الزمان، بصورة مباشرة أو غير مباشرة ويكفيهم فضلا أننا نعيش اليوم عصر الجينوم.

ولقد تراكمت المعارف خلال الخمسين سنة الماضية نحو ترسيخ مفهوم جديد للجين، فمع بدايات الخمسينات من القرن الماضى تم الإنجاز الأكبر فى علم الوراثة، بل قل فى علم البيولوجى عامة، وكشف الغطاء عن محتوى الصندوق الأسود (ألا وهو الجين)، وكان الفضل يرجع لكل من واتسون و كريك سنة ١٩٥٣فى التعرف على التركيب الجزيئي للجين، وهو الأحماض النووية خصوصا الـ DNA. منذ هذا التاريخ حتى يومنا هذا تضاعفت معارفنا عن تركيب (structure) ووظيفة (function) الجين، فمن مفهوم جين واحد – إنزيم واحد إلى جين واحد - سلسلة ببتيدية واحدة إلى مفهوم العقيدة الأساسية (central dogma) للتعبير الجيني التى تقر بأن الجين (gene) عبارة عن تتابعات من النيوكلوتيدات على طول الـ DNA تمثل إطارا (أو إطارات) للنسخ والترجمة (open reading frame) بالأضافة لعدة تتابعات عند النهايات ٣' و ٥' لاتنسخ ولكن تختص بالتحكم فى التعبير الجينى. ويتم نسخ الجين إلى نسخة من الـ RNA المرسال، وبمساعدة جزيئات أخرى من الـ RNA تترجم هذه النسخة إلى سلسلة ببتيدية (بروتين) محدد له وظيفة معينة. ظل هذا المفهوم للجين صامدا لأكثر من نصف قرن، ولكن مع نهاية القرن العشراون وبداية القرن الحادي والعشرون تجمعت لدينا حقائق جديدة جعلتنا نتشكك فى ثبات وصحة تلك العقيدة. ففى خلال السنوات القليلة الماضية، خصوصا مع مطلع القرن الحادى والعشرين، تجمعت أدلة ومشاهدات تجريبية تجعل من صحة وشمولية هذه العقيدة الأساسية لمفهوم الوراثة مشكوك فيها، وفيما يلى بعض من تلك الحقائق التى دفعتنا إلى إعادة التفكير:

١. التحقق من أن عدد الجينات المشفرة للبروتين (ORFs) فى الإنسان متدني عن المتوقع بشكل كبير (٣٠,٠٠٠ – ٤٠,٠٠٠ جين) بل تذهب بعض الدراسات الحديثة بعددها لأقل من ٢٥,٠٠٠ جين !! فكيف يفي الإنسان حاجته من البروتينات (حوالى ١٠٠,٠٠٠ بروتين) بهذا القدر الضئيل من الجينات، والذى قد يقارب ما هو موجود فى كائنات أقل تطورا !! وأن قدر الجينات المشفرة لا تشكل أكثر من ١,٥ – ٢ ٪ من جملة الجينوم فى الكائنات الراقية عموما.

٢. التعرف على عدد كبير (يزداد يوما بعد يوم) من الجينات غير المشفرة فيما يعرف بأسم ncRNAs، وهى التى تنسخ ولا تترجم إلى بروتين ولها العديد من الوظائف

التنظيمية بل والخلوية. واتضح أنها تمثل حوالى ٩٧ — ٩٨ ٪ من جملة النشاط النسخى (transcriptomea) فى الكائنات الراقية. وإذا ما اعتبرنا هذه التتابعات (ncRNAs) جينات !! فإنها ستزيد قدر المعلومات الوراثية من حدود الـ ٢٪ إلى حوالى الـ ٤٠ - ٥٠٪ من جملة الـ DNA الخلوى، لكن سيتبقى قدر كبير من الجينوم (حوالى ٥٠ ٪) غير محدد الوظيفة، فيما يسمى بالنفايات (Junk DNA) !! هل يمكن أن يكون هناك نفايات في النظام الحي ؟

٣. حالات الإعداد المغاير للمرسال الأولي (pre-mRNA alternative splicing) والتى تعددت صورها وأشكالها خلال الدراسات التى تمت خلال السنوات العشرين الماضية، وفيما يلي بعض منها:

(١) الإعداد المغاير من البداية (الطرف ٥') مثل جين (مجازا) الألفا أميلاز فى الفأر أو الأعداد المغاير من النهاية (الطرف ٣') مثل جين السلسلة الثقيلة للجلوبيين المناعي (immunoglobulin heavy chain) أو الأعداد المغاير من كلا الطرفين مثل جين تروبونين- ت فى العضلات (muscle troponin-T)، وكذلك حالات تحرير (تحوير) المرسال (mRNA editing) مثل جين apolipoprotein-B وغيرها.

(٢) الإعداد المتداخل لأطار واحد دون أدنى مشاركة بين المرسالات. Overlaping ORFs without shared coding sequences. وفى هذه الحالات ينسخ ويعد أحد الأطارات مشتملا الأكسون الأول مثلا فى الحالة الأولى ويستبعد فى الحالة الثانية لكونه يمثل أحد الأنترونات، ومثال هذه الحالات:

DNA complex *IP259/Dub80 in Drosophila melanogaster*.

(٣) الإعــداد النسـخى المشـترك لأثـنين مـن الإطـارات المسـتقلة. Co-transcriptional splicing between two ORFs مثـال حـالتي *P2Y11 & SSF1* فى الإنسان.

(٤) الإعداد النسخى المشترك المشتمل أحد الجينات الكاذبة. Co-transcriptional splicing with pseudogene.

(٥) ومثال ذلك:

CYP3A7 coding sequence and the pseudogene CYP3AP1 on human chromosome 7q21-q22.1

(٦) الإعداد النسخى المتبادل المتوازى (لكلا ذراعي الـ DNA لنفس الإطار).

ومثالها: Alternative trans-splicing

The modifier gene (mdg4) in D. melanogaster

وغيرها وغيرها من الحالات التى ينسخ ويعد فيها الإطار الواحد ليترجم لأكثر من بروتين (واللائحة كل يوم فى إزدياد)، ولكن أكثر الأمثلة الصارخة فى هذا الصدد فهى جينات العائلة الجينية المعروفة بأسم N-CAM حيث تكون مسئولة عن بناء طرز مختلفة من البروتين Neural Cell Adhesive Molecules، وهى أساسا مسئولة عن تكوين الذاكرة طويلة الأمد فى مخ الإنسان إلا أنها توجد فى كافة الأنسجة للتعارف والربط بين خلايا النسيج المتناظرة. المهم وجد أن أحد هذه الجينات يمكن أن يعد بطرق مختلفة قد ينتج عنها حوالى ١٠٠ طراز مختلف من البروتين iso-form !!

هـذه الحقـائق الجديـدة لم تتحـدى مفهـوم "الجـين الجزيئـي" وحسـب بـل أسـقطت الأعتقاد السائد منذ المندلية، بأن الجين وحـدة مسـتقلة محـددة لهـا ذاتيتهـا الخاصـة، فقـد أتضح أن الجين قد يشارك على المستوى الجزيئي فى أكثر من وظيفـة واحـدة، تركيبيـة أو تنظيمية. أن الهزة الكبيرة التى أصابت مفهوم الجين الجزيئى، والذى كنا نعتقد انه صامد إلى الأبد، جعلتنا نعيد التفكير فى مصداقية المسار الـذى نحن سـائرون فيـه (التبحـر فـى الدراسا ت الجزيئية عن الجينات). فمع تسليمنا المطلق بالأسـاس الجزيئـى للجينـات، إلا أن الواقـع يؤكـد أن هـذه الجزيئـات فـى نهايـة المطـاف هـى أدوات لتخـزين وتنـاول وتبـادل "المعلومات information " وأن لغـة التفـاهم هـى الشـفرة الوراثيـة (genetic code) مثلـها مثل أى لغة حية (مثـل الإنجليزيـة أو العربيـة...... أو غيرهـا). وبنـاء علـى هـذا التوجـه فيمكننا أن نعرف الجينات بأنها: وحدات معرفية – تحمل المعلومة لصفة ما ولكنها ليست مناطة تلقائيا بتنفيذ تلك المعلومة.

٢.١. المعلومات من تتابعات الـ DNA. DNA Sequence Information.

من المعروف لنا أن جزيئات الـ DNA هي الحامل للمعلومات الوراثية في الغالبية العظمى من الكائنات (الأستثناء في ذلك عدد قليل من الفيروسات حيث يحل محلها جزيئات من الـ RNA)، فهل تلك المعلومات محصورة فقط فيما نعرفه من معلومات العقيدة الوراثية الأساسية (الجينات المشفرة coding sequences) ؟ أم هناك معلومات أخرى خلافا لذلك ؟ لقد اتضح لنا أن هذه الجزيئات تحمل قدر من المعلومات أكبر مما كنا نتوقع، عرفنا بعضها ومازال أغلبها مجهولا. وفيما يلي ملخص لأهم المعلومات البيولوجية التي يمكن للـ DNA القيام بها أو التحكمبها:

- التتابعات المشفرة لأحماض أمينية (بروتين) وذلك إما :
    - بطريقة مباشرة مثل أغلب الجينات المعروفة mRNAs.
    - بطريقة غير مباشرة مثل حالات:
        - التجمع للجينات في البروتوزوا- تغير على مستوى الـDNA.
        - التحرير على مستوى الـRNA. RNA editing.
        - تفصيص الـRNA. RNA splicing.
        - تفصيص البروتيني. Protein splicing (الإنسولين).
- التتابعات غير المشفرة non-coding ، ومنها :
    - الناقل tRNA.
    - الريبوزي rRNA.
    - النووي الصغير snRNA.
    - التيلوميرازي telomeraseRNA.
    - وغيرها مثل RNAi
- التتابعات لبروتينات الأرتباط بمواقع محددة على الـ DNA protein binding sites.

- التتابعات الخاصة لتحديد الشكل الفراغى للجزئ، ومنها:
  - الألتفافات فى الحلزون intrinsic helix curvature.
  - تحديد مواقع النيوكلوسومات nucleosome positioning.
- تتابعات تحدد الثبات التركيبى الوظيفى، ومنها:
  - تحديد بداية النسخ transcription initiation.
  - نقطة أصل التضاعف origin of replication.
  - أماكن التطفر الساخنة "hot spots" mutational.

إذا ما كانت هذه بعض من الوظائف التى قد يقوم بها الـ DNA، فهل هناك أختلافات تركيبية لهذه الجزيئات قد يكون لها إنعكاسات وظيفية (قد لا نعلمها الآن ولكن سنعلمها فى المستقبل) ؟ وهل هذه الأختلافات التركيبية تعكس معلومات ما ؟ مثل هذا التساؤل يمكن فحصه من خلال عمل مسح scanning لتتابعات الـ DNA بحثا عن مناطق متشابهة مميزة بذاتها أو ما نطلق عليه المتكررات المتجاورة repeats. كذلك البحث عن أشكال مغايرة للتوزيعات الفراغية لجزيئات الـ DNA، أو أشكال الحلزون المختلفة الناجمة عن متكررات متجاورة من البيورينات أو البيرميدينات.

١.٢.١. المتكررات Repeats.

من دراسة تتابعات الـ DNA أتضح لنا وجود معدلات عالية من المتكررات على طول هذه الجزيئات يمكن تقسيمها إلى الآتى:

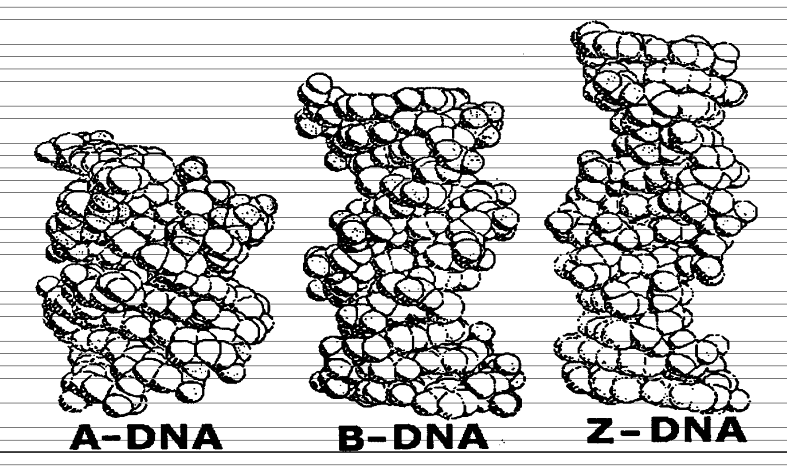
- متكررات مباشرة Direct repeats ⇨⇨ ، يوجد منها:
  - متكررات صغيرة متجاورة simple tandem repeats.
  - متكررات أكبر متجاورة longer tandem repeats
  - متكررات غير متجاورة non-tandem repeats.
- متكررات متواجهة Phased repears ⇦⇨، يوجد منها:
  - متكررات منعكسة inverted repeats ⇨⇦.
  - متكررات صورة فى المرآة mirror repeats ⇨⇦.
  - متكررات أرتدت reverted repeats ⇨⇨ .

وكل هذه الطرز من المتكررات repeats تنتشر فى جينومات كافة الكائنات التى درست حتى الآن بصورة كبيرة خصوصا فى مناطق السنترومرات centromeres والتيلوميرات telomeres الكروموسومية. وتتواجد بعض من صورها أثناء عمليات العبور الوراثى recombination (فى المناطق ثلاثية الأذرع – triplex) وكذلك أثناء التضاعف replication. هذه التراكيب قد يكون لها فائدة عند إجراء المقارنات بين الجينومات أو التتابعات المختلفة.

١. ٢. ٢. نماذج الحلزون DNA helix models.

لوحظ وجود أشكال مختلفة لحزون الـ DNA، ولكن أكثرها وجودا وتكرارا فى الكائنات الحية هو النموذج التقليدى الذى أقترحه واتسون وكريك سنة ١٩٥٣ والمعروف لنا جميعا، لذلك سمي فيما بعد بالنموذج البيولوجى biological-DNA ويعرف أختصارا بأسم B-DNA. بالأضافة للنموذج البيولوجى وجدت نماذج أخرى مختلفة وتوجد فى الأنظمة الحية ولكن بندرة عرف أهما بأسم نموذج A-DNA ونموذج Z-DNA. وشكل (١-٢) يوضح هذه النماذج الثلاث كما نشرها Dickerson et al. سنة ١٩٨٢ لقطع متساوية لكل منها.

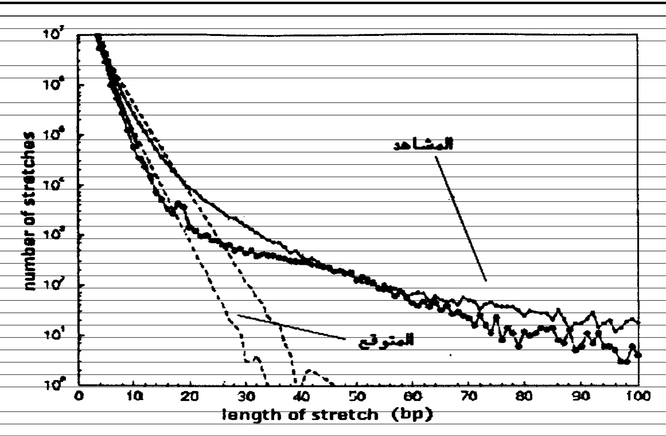
عائلة الحلزونات من الطراز A-DNA أكثر تواجدا فى جزيئات الـ RNA مذدوج الذراع ds-RNA أو الهجن بين الـ DNA والـ RNA (RNA/DNA hybrids). ويمكن أن يتواجد من خلال بعض التتابعات حيث تتكرر تتابعات متجاورة (خمس نيوكلوتيدات) من البيورينات purines . اما عائلة الـ B-DNA فهى الغالبة فى أنوية الكائنات الحية، ولكنها قد تظهر قدرا من الأختلافات، فمن المعروف أن ١٠ أزواج من النيوكلوتيدات المقترنة تمثل لفة واحدة للحلزون المزدوج، لكن وجدت تتابعات محدد قد تؤدى لتكوين اللفة مابين ٩ و ١٢ نيوكلوتيدة مقترنة. العائلة الثالثة هى عائلة Z-DNA وهى أقل الثلاث حدوثا، ولكن عند تتابعات معينة مثل متكررات GC (GCGCGC)



شكل (١-٢): رسم توضيحي لنماذج الـ DNA (A-DNA, B-DNA, Z-DNA)، كل ممثل بقطعة طولها ١٢ زوج من النيوكلوتيدات المقترنة.

يتكون الحلزون Z-DNA عادة في الكائنات حقيقية النواة eukaryotes. كذلك لوحظ أن بعض الجزر من CpG العادية أو المضاف لها مجاميع المثيل تكون هذا الحلزون في التجارب المعملية، كذلك عزلت بعض البروتينات التي تفضل الارتباط بمناطق الحلزون Z-DNA.

الملفت للنظر أن الدراسات الحديثة أوضحت أن حدوث وتكرارات هذه الحالات في البكتيريا أعتباطي ويرجع للصدفة ولكن وجد أن تكرار هذه الحالات غير أعتباطي وينحرف معنويا عن المتوقع في دراسة على جينوم الإنسان، كما هو موضح بشكل (١-٣).



شكل (١-٣): علاقة بيانية بين تكرارات حالات الحلزون الثلاثة وأطوالها كما هو متوقع وكما هو مشاهد لتتابعات كروموسوم ١ فى الإنسان.

حقيقة كون أن تكرار هذه الحالات فى الكائنات حقيقية النواة لا يرجع للصدفة، بينما الحال مختلف فى الكئنات غير حقيقية النواة حيث وجد أن أغلبها يرجع للصدفة، هذه النتائج تلفت النظر إلى أهمية الصدفة فى دراستنا للأنظمة البيولوجية. لذلك وجب التعرف على مفهوم الصدفة ولم بصورة مبسطة.

١. ٢. ٣. التركيب الثانوى لجزيئات الـ RNA.

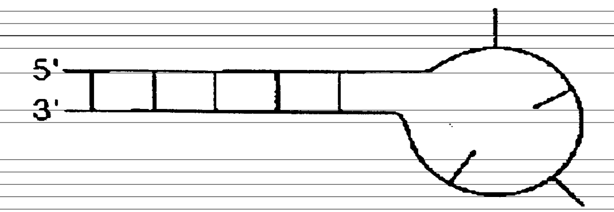
RNA Secondary Structure.

جزيئات الـ RNA على إختلاف أشكالها وطرزها تختلف عن جزيئات الـ DNA وإن كانت لاتقل عنها أهمية من حيث دورها فى نقل وتداول المعلومات الوراثية. جزيئات الـ RNA تختلف عن الـ DNA فى كونها جزيئات خطية (مفردة الذراع) من تجمع وحداتها.

الأقترانات التقليدية ما بين الجوانين والسيتوسين والأدانين واليوراسيل، ولكن هناك أيضا أقترانات غير تقليدية أكثرها شيوعا ما بين الجوانين واليوراسيل. هذه الأوضاع والأشكال الجزيئية الناجمة عن اقترانات تؤدى إلى ما يسمي بالتركيب الثانوى secondary structure وهو حالة وسطية بين التركيب الخطى والتركيب ثلاثى الأبعاد

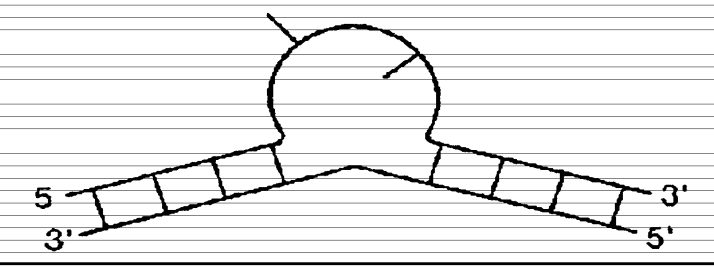
3D structure. وبناء على التركيب الثانوى تتكون عدة تراكيب جزيئية مميزة يمكن إجمالها فى الحالات التالية:

- الأنتفاخ القاعدى stem loop أو انتفاخ دبوس الشعر hairpin loop، كما هو مبين بالرسم التالى:

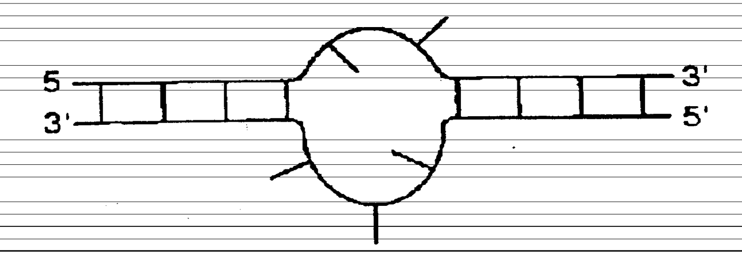


وعادة لا يقل الأنتفاخ عن أربعة نيوكلوتيدات غير مقترنة.

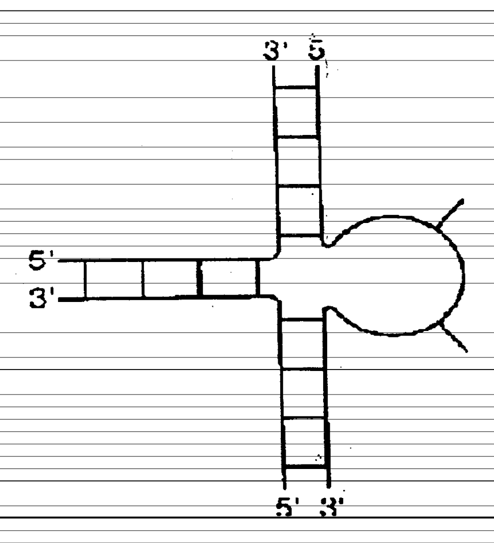
- الأنتفاخ البرميلى Bulge loop، ويتكون عادة عند مناطق عدم الأقتران فى طرف واحد من الإنتفاخ، كما هومبين بالرسم التالى:



- إنتفاخ داخلى interior loop يتكون عند مناطق عدم الإقتران على طرفى الإنتفاخ، كما هو مبين كالآتى:

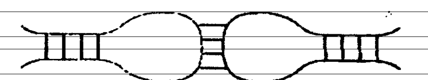


- إنتفاخ الوصلة junction loop، ويتكون من أقتران جزيئين من الـ RNA مكونـان تراكيب ثانوية مشتركة، كما هو مبين كالتالي:

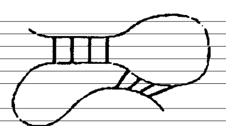


- التراكيب الثلاثية tertiary structures وهى تنتج من الأندماج بين أكثر من جزئ، ومنها أشكال عدة يمكن تمثيل أهمها كالتالى:

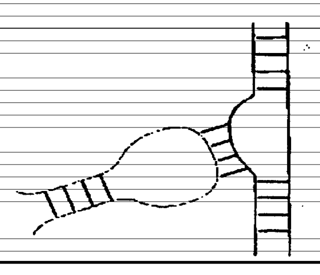
- إنتفاخات متقابلة kissing loops



  - العقد الكاذبة psuedoknots



  - التداخل بين إنتفاخ برميلى ودبوس الشعر hairpin and bulge interaction



١.٢. المعلومات والصدفة و التحكم. Information & Chance & Control.

نحن نتعايش فى كل لحظة من حياتنا مع عدد من الأحداث غير المتوقعة (unexpected) أو غير المتحكم فيها (uncontrolled) والتى يمكن أن تندرج تحت ما تعارفنا عليه بأسم الصدفة (chance) ، وهى تحدث دوما فى العالم من حولنا فهى جزء لا يتجزأ من ظواهر الكون. فهل تكرار الصدف من حولنا دفعنا يوما للتفكير فى طبيعة

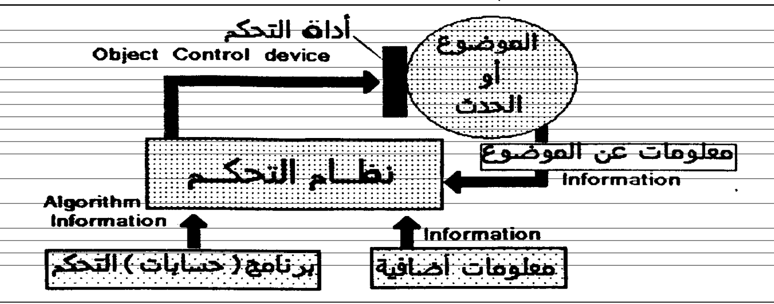
هذه الصدف ؟ فى أغلب الأحيان لانلتفت كثيرا لتلك الصدف من حولنا، وهذا ليس نوعا من الأنكار لوجودها ولكن لتآ لفنا معها واعتبارها أمر مفروغ منه. و لكن وبدون أدنى شك لقد حاول البشر معرفة لماذا يحدث غير المتوقع من الصدف، بل يمكننا القول بأن الحضارة الأنسانية منذ فجر التاريخ وحتى الآن هى فى سعى دؤوب لمعرفة غير المتوقع (الصدف) أو المجهول. فيمكننا القول بأن الصدفة هى مرادف عدم المعرفة أو الجهل، فكلما زادت معلوماتنا كلما تضائلت فرص الصدفة من حولنا. لتوضيح الأمر أكثر دعنا نسوق المثال التالى : اذا ما كنت فى الخارج راجعا إلى منزلك ظهرا بعد دوام العمل كما هى عادتك يوميا – وضغطت على جرس الباب الخارجى لمنزلك – فإنك عادة تكون متأكدا بقدر كبير بأن احدا (زوجتك مثلا) سيفتح لك الباب، ولكن فى يوم من الأيام لم يفتح لك أحد قل بعد عشر دقائق، هذا الأمر غير المتوقع (الصدفة) سيصيبك بالدهشة بل بالإنزعاج لأن احدا ليس بالمنزل – ولكن إذا ما شككت فى إنقطاع التيار الكهربى عن شقتك – وبدات بالطرق بيديك ولم يفتح لك أحد – سيزداد إنزعاجك من المصادفة خصوصا إذا اتضح لك أن نور السلم يعمل وتزداد حيرتك من وقع المصادفة (كل ذلك راجع لجهلك بالأمور الجارية). فى اثناء ذلك هداك تفكيرك أن تتصل بزوجتك على "المحمول" فتخبرك بأنها اضطرت لترك المنزل لطارئ ما – هنا تطمئن وينتهى تأثير الحادثة عليك وتتقبل الأمر ببساطة. إذن انتهى تأثير الصدفة (الأحداث غير المتوقعة) عليك بمجرد معرفتك بحقيقة الأمر، اى أن المعرفة تقلل من أحتمال الصدفة من حولنا. كذلك مع توفر وسيلة للتحكم control قللت من أحتمال حدوث الصدف. ففى مثالنا السابق – صادف هذا الرجل حدث غير متوقع (صدفة) كادت أن تخرجه عن إتزانه، ولكنه تغلب عليها، فلماذا ؟ لأنه حاول التحكم فى الحدث مستعملا عقله كأدات للتحكم، ثم إستعان بوسائل متاحة لزيادة معلوماته (الأتصال بالمحمول أو إنارة نور السلم وغيرها....).

دعنا الآن نخرج من حدود هذا المثال الضيق، لأستخلاص بعض القواعد العامة، فتبعا لقانون الديناميكا الحرارية الثانى: ففى الأنظمة المغلقة، فإن النظام يصل إلى حالة من استنزاف للطاقة تسمي " الأنتروبيا" (entropy) بعدها ينعدم أو يفنى النظام. أما نظرية التطور (على الجانب الآخر) فتقر بتعقيد النظم الحيوية وانتظامها الدقيق بل الفائق وتسلسلها عبر الزمن من منشأ واحد. هذا التعارض البين بين كلا النظريتان حول

مصير الأنظمة الحية، من الفناء المحتوم تبعا لقوانين الديناميكا الحرارية - اوالبقاء والتعقيد المتحكم فيه تبعا لنظرية التطور. هذا التعارض دفعنا للتفكير لنستخلص من تلك الظواهر والحقائق المادية ، ملاحظة غاية فى الأهمية وهى: "أن التحكم (control) و توفير المعلومات (information) فى الأنظمة عامة (والبيولوجية خاصة) يؤدى إلى تقليل الأنتروبيا ويسمح بقدر من الإنتظام (order)". إذن التحكم والمعرفة هما وسيلتنا لتقليل الصدف. شكل (١-٤) يلخص وسائل تقليل الصدف فى الأنظمة المختلفة (مادية كانت او حيوية). فأذا ما تعاملنا مع حدث (موضوع) ما بغرض تقليل الأنتروبيا (تقليل الصدف)

فلابد من وجود اداة للتحكم (مثل الثرموستات فى جهاز التكيف يعمل عند درجة حرارة محددة)، لكن لابد من وجود عقل للتحكم قد يمثله برنامج حسابى algorithm يتحكم فى تلك الأداة. لكن لكى يعمل هذا البرنامج لابد من مده بالمعلومات، أولا عن الغرض من ألتحكم (ضبط درجة حرارة الغرفة) وكذلك معلومات أضافية (عن ظروف الجو خارج الغرفة مثلا وغيرها وغيرها). إذا توفر ذلك فستكون لنا فرصة كبيرة للتغلب عن الصدف غير المتوقعة (تقلب الطقس مثلا).

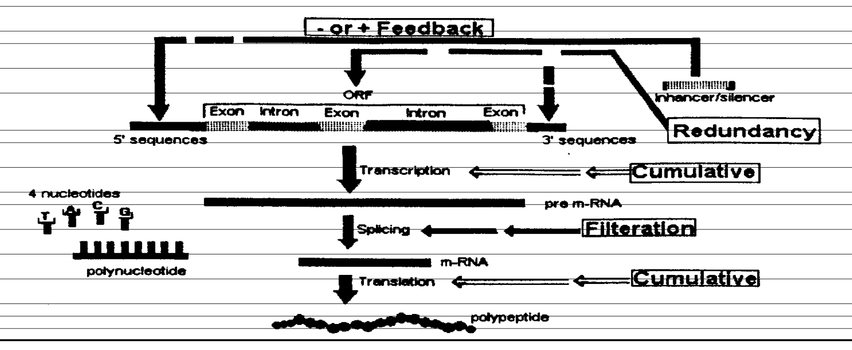
مما سبق يتضح لنا أن حرب البشر المستمرة على غير المتوقع من الصدف (أو عدم الإنتظام - disorder) كان ومازال لها جبهتان رئيسيتان للقتال، الأولى مباشرة للقضاء على الصدفة (الشوشرة) العشوائية وقد أثبت التاريخ إستحالة نجاح هذه الجبهة، أما الجبهة الثانية فهى الدبلوماسية (إذا جاز التعبير) بمعنى التعايش مع الصدف (التداخل العشوائى) ومحاولة التقليل من تأثيراتها دون إنكارها أو محاولة التقليل من قدرها، ويمكن تمثيل الأستراتيجية الثانية بحالة المحادثة فى الهاتف فى ظل وجود تداخل بالخط (شوشرة) فلا يمكننا التخلص منها أثناء المحادثة ولكن لنتغلب على الوضع يمكن أن نرفع أصواتنا او نكرر ما نقول عدة مرات، وقد أثبتت الأيام أن الجبهة الثانية هى المتاحة لنا فى حربنا على غير المتوقع من الأحداث.



شكل (١-٤) : رسم تخطيطى عن وسائل تقليل غير المتوقع (الصدفة) لحدث ما.

ففى حربنا الدبلوماسية على الصدفة فإن وسائلنا عديدة مثل التكرار (redundancy) للمعلومات و تراكمها (cumulation) وغربلة (filtiration) الصالح منها مع توفر أنظمة للتحكم (control). وهذه الوسائل نجدها ممثلة وبوضوح فى أى نظام وراثى، كما هو مبين بشكل (١-٥).

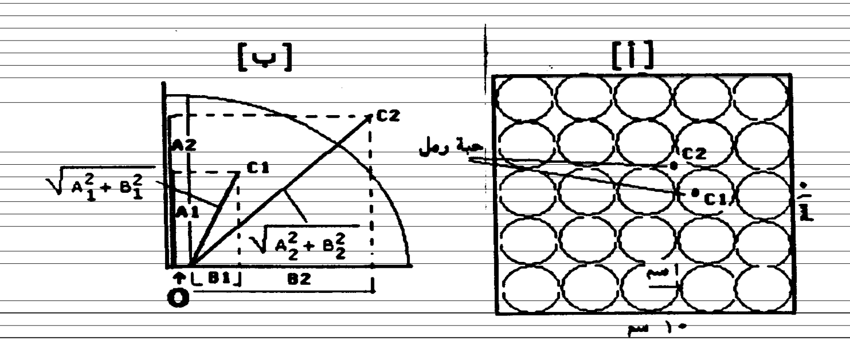
ففى هذا النظام توجد أنظمة متعددة موجبة أو سالبة للتحكم (سبق دراسة عدد كبير منها فى مقررات وراثية أخرى، ولا يسع المجال هنا لتكرارها). وكذلك نلاحظ التكرار للمعلومة الوراثية الواحدة (هناك نسخ متكررة من الجين الواحد)، كذلك فإن تراكم وتجميع المعلومة الوراثية عدة مرات ظاهرة عامة، ينسخ الجين الواحد ويترجم عدة مرات قد تصل مئات المرات للتأكد من وصول المعلومة المطلوبة. وظاهرة الأنتقاء للأكسونات دون الأنترونات خلال عمليات الأعداد والتفصيص للمرسال الأولى ماهى إلا وسيلة للغربلة والتصفية للتخلص من غير المرغوب من المعلومات مع الأبقاء على المرغوب منها.



شكل (١-٥) : رسم توضيحى لنظام جينى عام يوضح نقاط التحكم وتقليل الشوشرة المتاحة.

مما سبق وضح لنا أهمية التعايش والقبول بوجود الصدفة فى كافة مناحى الحياة من حولنا، ثم وجب علينا بعد ذلك محاولة التقليل من تأثيراتها ما أمكن. فى محاولتنا هذه نحن أشبه بمن يلعب لعبة من لعب التسلية (a game) فعلينا أن نحدد خياراتنا بين أفعل ولاتفعل وعلى إختيارى هذا ستتوقف النتيجة بين الفوز والخسارة، ومن هنا ظهرت "نظرية اللعبة" (Game theory) التى أصبح لها تطبيقات واسعة خصوصا فى علوم الأقتصاد والأجتماع والسياسة. أذن علينا حساب القرار الصائب، وهذه الطريقة (وغيرها من الطرق الحسابية والأحصائية) عليها أن تزيد من إحتمالات النجاح وتقليل إحتمالات الفشل عن طريق حساب الأحتمال (probability) عند مستوى مصداقية محدد. فنحن فى حياتنا نتعامل من الأحتمالات (الحدس المقبول) بصورة مستمرة دون أن نلتفت لذلك، فعلى سبيل المثال – إذا ما سألت عن مساحة الدائرة – فستكون أجابتك المباشرة مساحة الدائرة تساوى مربع نصف القطر مضروبا فى π ‏(r^2 π) فإذا كان لديك دائرة نصف قطرها = ١سم، فإن مساحتها = ١^٢ × ٣,١٤ = ٣,١٤ سم^٢. فمساحة الدائرة تقريبية أذن حيث أن π هنا قيمة مقربة، فهل خطر فى ذهنك يوم هل يمكن أن احسب قيمة π الحقيقية ؟ وهل يمكننى ذلك ؟ هنا يأتى دور طرق حساب الأحتمالات الحسابية، ولتبسيط الفكرة اكثر، دعنا نستعمل أحد الطرق الشهيرة وهى طريقة (Monte Carlo method) نسبتا للمدينة الشهيرة بنوادى القمار وخصوصا لعبة الروليت التى بنيت عليها هذه الطريقة.

ولكن قبل ذلك دعنا نحل المسألة تجريبيا، فلو أحضرنا قطعة من ورق الكارتون أبعادها ١٠ × ١٠ سم أي مساحتها ١٠٠سم^2 ورسمنا عليها دوائر كل نصف قطرها ١سم بحيث تكون متماسة تماما (كما هو موضح في شكل ١-٦أ) فسيكون عدده ٢٥ دائرة، ثم أتينا بحبة رمل صغيرة وألقيناها على الرقعة عدد كبير من المرات، ثم حسبنا عدد المرات التي وقعت فيها حبة الرمل داخل دائرة من الدوائر (C_1) وعدد المرات التي وقعت فيها بالمسافات البينية بين الدوائر (C_2)، تجدر الملاحظة أننا في هذه التجربة تمثلنا لعبة الروليت الشهيرة. فقل كانت نتيجة ألقاء حبة الرمل ١٠٠٠ مرة - ٧٠٠ وقعت داخل أحد الدوائر و٣٠٠ مرة في المسافات البينية، فيمكن القول أن مساحة الدوائر بالرقعة تمثل ٧٠٠/١٠٠٠ = ٧/١٠، أي يمكن حساب مساحة الدوائر ١٠٠ × ٧/١٠ = ٧٠سم^2 وحيث أن عدد الدوائر الكلي = ٢٥ أذن يمكن حساب مساحة الدائرة الواحدة ومنها يمكن حساب قيمة π، والتي ستساوي في هذه الحالة ٢,٨.



شكل (١-٦) : محاولات لحساب قيمة π (أ) تجريبيا و (ب) حسابيا.

اتضح لنا مما سبق أنه يمكننا أن نحسب تجريبيا قيمة π حسب الأحتمال المحسوب، فهل يمكن وضع نموذج حسابي (mathematical model) لتقدير الأحتمال دون اللجوء للتجربة خصوصا أن مثل هذه التجارب عرضة للأخطاء التجريبية العديدة (فعلى سبيل المثال عدم التأكد من مكان حبة الرمل إذا ما وقعت على الحدود وكذلك احتمال عدم إستواء الرقعة بنفس الدرجة وغيرها وغيرها من العوامل غير المتحكم فيها). لعمل هذا

النموذج الحسابى افترض اختيار رقمين عشوائيا مثل A و B بحيث أن يكونا أصغر من الواحد الصحيح وأكبر من الصفر ($0<N<1$) و للتحقق من وقوعهما داخل أو خارج أحد الدوائر دعنا نتصور ربع أحد تلك الدوائر، كما هو موضح بشكل(١-٦ب). فمن النقطة O اختار رقمين عشوائيا A_1 و B_1 وحدد نقطة تلاقيهما C_1. أعد المحاولة وحدد النقاط A_2 و B_2 و C_2. فالنقطتان C_1 و C_2 مواقع محتملة لحبة الرمل فى التجربة السابقة وتبعا للعلاقات الحسابية يمكن التوصل للمعادلتين التالتين: إذا ما كان $A^2 + B^2 \leq 1$ فإن الرقم سيقع داخل الدائرة لكن لو كان $A^2 + B^2 > 1$ فإن الرقم سيقع خارج الدائرة. اذن يمكننا الآن حساب احتمال كلا الحدثين دون رقعة ولا حبة الرمل ولا التكرار فى رمي الحبة. بدون شك أن النموذج الحسابى أفضل، فبه نتخلص من الكثير من الأخطاء التجريبية التى تؤثر على صحة التقدير، ولكن مازال علينا أن نختار تلك الأرقام العشوائية وحساب المعادلات لكل منها مئات المرات مما يضعنا أمام عمل شاق قد تفوق صعوبته رمي حبة الرمل العديد من المرات، أى لم نوفر شيئا فى المجهود والوقت. هنا جاء دور الكمبيوتر فهو يمكن أن يولد عدد لامتناهى من الأرقام العشوائية وحساب نتائج المعادلات فى دقائق بل فى ثوانى فى بعض الموديلات الحديثة. الأعتماد على الكمبيوتر إذن ضرورى ولكن الكمبيوتر لا يمكنه القيام بتلك المهام من تلقاء نفسه بل يجب كتابة تلك الأوامر والحسابات فى صورة برنامج رياضى يوضح للكمبيوتر مهامه المطلوبة منه، هذا البرنامج لحساب احتمالات الأحداث هو ما تعارفنا علية بأسم طريقة مونت كارلو فى المثال السابق.

١. ٤. الشبكات الجينية Gene networks.

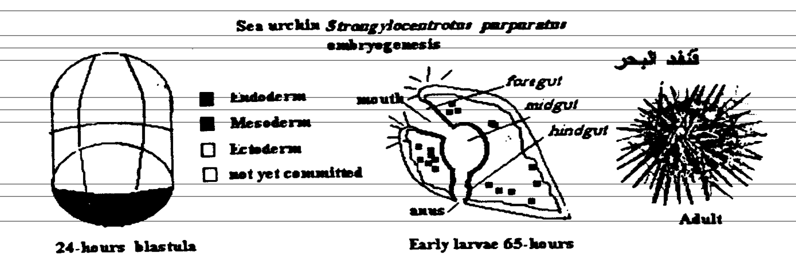
لقد عايشنا خلال أواخر القرن الماضى وأوائل القرن الحالى بذوغ عصر الجينومكس والمعلوماتية الحيوية وإنجازاتهما المهولة فى مجال تعرفنا على طبيعة المعلومات الوراثية، ولكن أدى بنا هذا الحال (زيادة المعلومات وتعقدها) إلى التخبط فى فهمنا للمنظومات الجينومية وكيفية تفاعلها وتداخلها مع بعضها البعض للوصول للشكل المظهرى. هذا الوضع دفع عدد من علماء الوراثة للقول بأن القرن الحادى والعشرين هو عصر ما بعد الجينومكس post genomic area، ففى تعليق لأحد أشهر الوراثيين وهو Lee Hartwell

(الحائز على جائزة نوبل سنة ٢٠٠١ لأكتشافه بروتينات الـ cyclins ودورها فى دورة إنقسام الخلية) قال أن طلاب الوراثة فى المستقبل سيكون همهم التعرف على الدوائر والشبكات الجينية gene circuits & networks عوضا عن دراسة الجينات بذاتها، و يجب علينا أن نتجه نحو بناء النماذج الوراثية أو ما يسمى بأنظمة المحاكاة simulation systems فى مجال الوراثة، فلو توفر لنا نظام يحاكى الخطوات الوراثية للأصابة بالسرطان مثلا، فسنكون أكثر قدرة على تفهم هذا المرض بل سنكون قادرين على تصميم العقاقير المناسبة لكل حالة بحالتها.

أن محاولات الباحثين لبناء الشبكات الجينية لحالات بعينها مازالت فى مراحلها الأولى بل أن الإنجازات المتاحة فى هذا المجال نادرة ومازالت فى مراحلها الأولى وبعيده عن الوضوح والأكتمال، ولكن فى سبيل بناء تلك الشبكات والدوائر الجينية يجب أن لا ننسى أنها نماذج نظرية وتختلف عن الأنظمة الحية، وهدفنا من بناء تلك النماذج سيكون فقط لزيادة قدرتنا على التنبؤ بالمخرجات outputs المتوقعة من تلك المدخلات inputs الجينية. وفى محاولتنا لتفهم هذا الموضوع سنكتفى بمناقشة أحد الدوائر لجين واحد فقط، لتجنب التعقيدات المهولة التى قد تقابلنا فى هذا المجال.

١. ٤. ١. الجين endo16 ومراحل تكوين جنين قنفذ البحر.

أختص عالم الوراثة الشهير Eric Davidsons وتلاميذه ومساعديه فى معهد كاليفورنيا للتكنولوجيا بدراسة هذا الجين من أوائل تسعينات القرن الماضى وحتى الآن، حيث أجروا آلاف التجارب لعشرات السنين للوصول لتفهم هذا الجين، والذى يعتبر أكثر الجينات دراسة حتى الآن. قنفذ البحر (الريتزا) sea urchin, *Strongylocentrotus purpuratus* معروف فى تجارب التشكل والتكوين، فمراحل تشكله الجنينى معروفة بالتفصيل (الأجنة شفافة يسهل تتبعها ميكروسكوبيا). فبعد الإخصاب تنقسم الخلايا الجنينية مكونة عدة طبقات تعرف بأسم ectoderm, mesoderm, endoderm وتشكل مرحلة البلاستيولا blastula من حوالى ٥٠٠ خلية بعد حوالى ٢٤ ساعة، بعد هذه المرحلة يتبعها مرحلة الجاستريولا gastrula حيث يتم إنتظام خلايا طبقة الـ endoderm للداخل لتكون جدران الأحشاء الداخلية لليرقة وذلك فى حوالى ٦٥ ساعة لتكون اليرقة، كما هو موضح بشكل (١-٧).

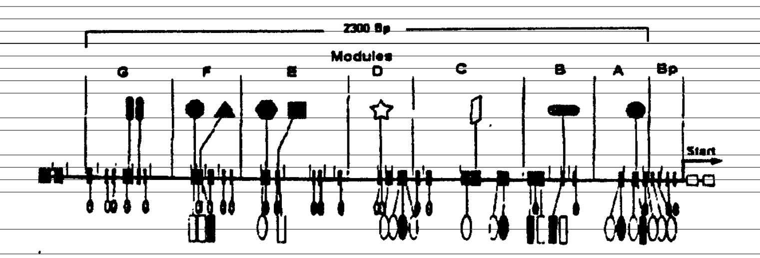


شكل (١-٧) : مراحل التكون الجنينى فى حيوان قنفد البحر *S. purpuratus*.

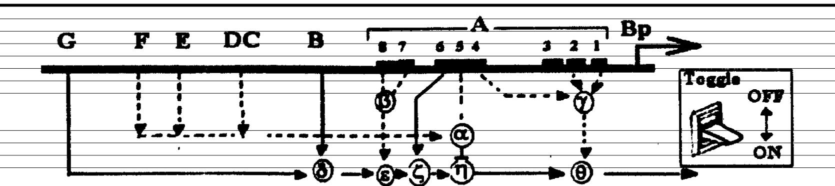
ومن دراسات Davidsons *et al.* أمكن التعرف على الجين endo16 وهو الجين المسئول خلال مراحل الـ gastrula على تشكل خلايا الـ endoderm لتغلف الأحشاء الداخلية وان نشاطه يحدد زمنيا ومكانيا فى هذا النوع من الخلايا دون غيرها لتشكل منظومة تحكمية غاية فى الدقة والتحكم. فقد وجد أن مناطق البرومتر البعيدة (لمزيد من التفصيل يمكن الرجوع إلى المتينى - ١٩٩٧) تمتد على طول المناطق السابقة upstream لهذا الجين بطول حوالى ٢٠٠٠ زوج من القواعد المقترنة فى المناطق البين جينية intergenic أو ما كان يعرف بأسم النفاية junk DNA ! (بدون وجه حق) وهى مقسمة إلى ثمانية مناطق modules متتالية من الـ cis-regulatory DNA منها البروموتر الأساسى (القريب) حيث يرتبط أنزيم النسخ pol II و المميز بالحروف (Bp)،والمناطق الأخرى المميزة بالحروف من A إلى G كل منها حوالى 300 bp وكل به أكثر من ١٠ مواضع حيث ترتبط العديد من البروتينات المنظمة للنسخ (المنشطة والكابتة)، كما هو مبين بشكل (١-٨).

بأستعمال أساليب الهندسة الوراثية لمعرفة وظيفة كل مقطع module من تلك المقاطع على حدة، أتضح أن المقاطع G, B, A تنشط أو تنبه نسخ هذا الجين فى خلايا الـ endoderm وخصوصا فى منطقة الـ midgut من المراحل المبكرة من التشكل، وان المقاطع F, E تكبت عمله فى خلايا الـ ectoderm بينما المقاطع D, C تكبت عمله فى

خلايا الـ mesoderm. وبتراكم المعلومات بهذا الخصوص اتضح أن المقطع A هو المقطع الأساسى القادر على تنبيه منطقة البروموتر الأساسى Bp، وان باقى المقاطع تظهر تأثيرها من خلال الأرتباط بتتابعات محددة داخل A ولا يمكنها الأرتباط بالبروموتر الأساسى مباشرة. هذه المعلومات الشيقة دفعت Davidson وفريقه إلى تصور نظام التحكم لهذا الجين من خلال تبنى مفاهيم وأفكار الهندسة الكهربائية فى تصميم الدوائر الكهربائية circuits، ولكن لم يكن هدفهم بطبيعة الحال تصميم دوائر من الأسلاك والموصلات الكهربائية، فهناك فرق جوهرى حيث نتعامل هنا مع أنظمة حية. وشكل (١-٩) يمثل الدائرة الجينية endo16.

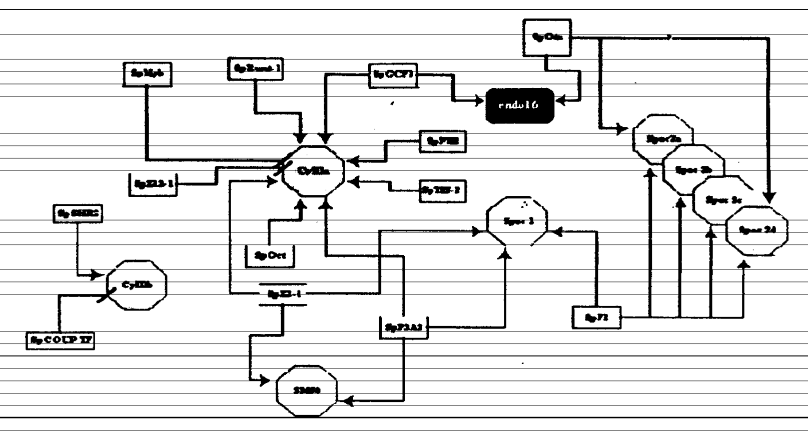


شكل (١-٨) : مناطق البرومتر البعيدة لجين endo16 مميزة لعدة مناطق حيث يمكن ان يرتبط بها البروتينات المنظمة للنسخ (الأشكال الهندسية المختلفة).



شكل (١-٩) : رسم توضيحى لدائرة الجين endo16، كما اقترحها Davidson ومعاونيه (فى اقصى اليمين شكل يمثل مفتاح toggle).

وفى هذه الدائرة يمكن تصور المقاطع من A إلى G كمصادر للأشارات signals كل يخرج منه موصلات، وعند إلتقاء إشارتين مع بعضهما فتمثل على هيئة عقدة node (ممثله فى الشكل بدوائر داخلها حروف يونانية) ويمكن تصور وجود ترانسيستور صغير عند كل عقدة عليه القيام بالحسابات لأتخاذ القرار. فعند العقدة α يتم تلقى الإشارات من المقاطع F, E, C, D حيث ترتبط بالتتابع رقم ٥ على المقطع A، وعند العقدة β تلتقى الإشارات من التتابعين رقمي ٧ و ٨ على A لتعظيم نشاط المقطع B، وعند العقدة γ تلتقى الأشارات من التتابعات رقمى ١ و ٢ لتعظيم الإشارات من المقطع G، وعند العقدة δ تلتقى الإشارات الموجبة القادمة من المقطعين B وG وهذه العقدة متوقفة على زمن العمل، والعقدة ε تتلقى الإشارات من العقد β و δ وتجمعهما وتصدر إشارة موحدة من مجموعهما. اما عند العقدة ζ، فيجب تصورها كمفتاح toggle يسمح إما بالعمل ON او القفل OFF (مثل مفتاح الكهرباء مثلا) فهو يتلقى إشارات من التتابع رقم ٦ على A (المعلومات الخاصة بالأرتباط بالبروموتر الأساسى Bp لبدأ النسخ) وكذلك يتلقى الإشارات القادمة من ε الذى يعكس معلومات المراحل التكوينية وزمنها. هنا المفتاح زيتا ζ هو الذى يسمح بالعمل للجين، ولكن لتحديد نوع الخلايا التى يتم فيها العمل دون غيرها وجب وجود العقدة η حيث تتجمع الإشارات الكابتة من كافة المقاطع المختصة لوقف العمل فى خلايا ectoderm و خلايا mesoderm والتى يحددها موقعهما فى النسيج الجنينى. وفى حالة السماح بالعمل (خلايا الـ endoderm) تنقل الإشارات إلى θ حيث تتلقى إشارات α المساعدة على الأرتباط مع Bp. ويمكننا أن تصور قدر المجهود والوقت الذى بزلته هذه المجموعة من الباحثين الأكفاء للتوصل لهذا النموذج الفريد لدائرة الجين endo16 وهو يمثل وحدة واحدة من منظومة جينومية قد يبلغ أعضائها الآلاف من الجينات، إلا أن هذا الوضع لم يثنى Davidson وفريقه سنة ٢٠٠٥ فى وضع أول تصور لشبكه جينية gene network تمثل مراحل التشكل الأولى لقنفد البحر وتشمل أكثر من حوالى ٣٠ جين مختلف، كما هو مبين بشكل (١-١٠). ومع أن هذه الشبكة مبدئية ويشوبها الكثير من القصور إلا ان المستقبل يبشر بنتائج طيبة فى هذا المجال.



شكل (١- ١٠) : رسم تخطيطى للشبكة الجينية لمراحل التشكل المبكرة فى قنفد البحر

(المربع الأسود يمثل الجين endo16 فى هذه الشبكة).

٢. الجينومكس التركيبي

Structural Genomics

إعداد: أحمد المتيـنى

لكي نقرب إلى الذهن مدى الجهد والصعاب التى قابلها المختصين لإنجاز الدراسات والمشاريع لسلسلة أو فك تتابعات جينومات بعض الكائنات والتى انتهت، خصوصا فى مراحلها الأولى، يمكن أن نجرى المقارنة التالية، فلو اخذنا مثلا الفيروس (البكتريوفاج لامدا) فجينومه حوالى ٥٠٠٠٠ bp فإذا افترضنا أن الصفحة المكتوبة بخط صغير جدا تحتوى على تتابعات قدرها ٢٥٠٠٠ bp ، فأن جينومه بالتالى سيملأ صفحتين. إذا انتقلنا إلى *البكتيريا (E. coli)* فإن تتابعاتها ستشغل كتيبا صغيرا من ٢٠٠ صفحة، فإذا انتقلنا إلى كائن دقيق حقيقى النواة مثل الخميرة فأن تتابعات جينومه ستملأ مجلدا صغيرا من حوالى ٥٠٠ صفحة. أما الدودة *الثعبانية (C. elegans)* والنبات الصغير (*A. thaliana*) فلهما جينومين متساوين تقريبا فى الحجم، فجينوم كل منهما سيشغل ثلاث مجلدات كبيرة، أما إذا وصلنا للإنسان فإن جينومه سيشغل حوالى ٨٠ مجلدا كبيرا. والجدول التالى يلخص قدر أزواج النيوكلوتيدات فى مجاميع الكائنات الرئيسية.

مجموعة الكائنات	حجم الجينوم (bp)
الفيروسات	300 – 350,000
غير حقيقية النواة	250,000 – 15,000,000
حقيقية النواة – وحيدة الخلية	12,000,000 – 50,000,000,000
حقيقية النواة – عديدة الخلايا	20,000,000 – 600,000,000,000

وبزيادة المعلومات لعلم الجينومكس وتشعبها اتفق العلماء على تقسيمه إلى عدة مجالات وهى :

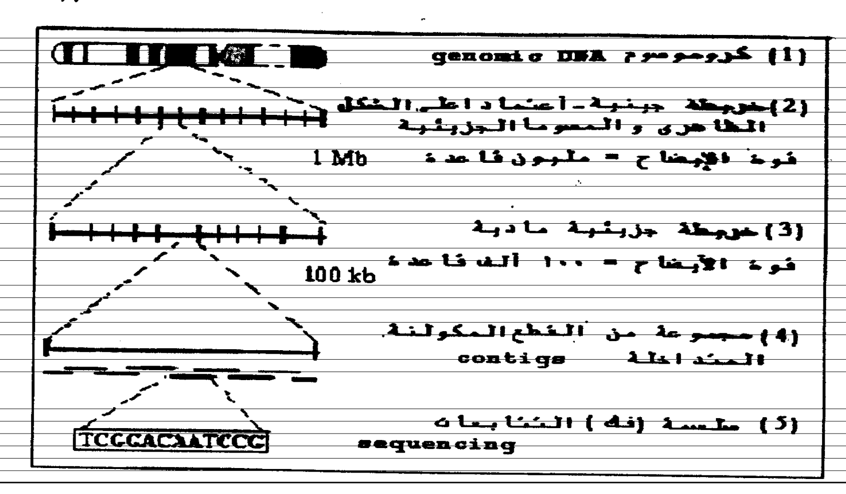
(١) الجينومكس التركيبي (Structural genomics) الذى يختص بدراسة التركيب الجزيئي (المادي) لجينوم ما ،

(٢) الجينومكس الوظيفى (Functional genomics) الذى يختص بالتعرف على النشاط أو الفعل لجينوم ما،

(٣) الجينومكس المقارن (Comparative genomics) ويختص بمقارنة جينومات الأنواع المختلفة بعضها ببعض.

٢.١. مدخل Introduction

عادة يحدد حجم جينوم الكائن مدى الجهد والصعوبة المتوقعة لدراسته، لكن يجب ان نوضح بأن حجم الجينوم لكائن ما لا يعكس بالضرورة مدى تطور هذا الكائن فيما يعرف بأسم معضلة حجم الجينوم (C-paradox) فبعض البرمائيات وبعض النباتات تتميز بجينوم (C-value) أكبر من ذلك فى الإنسان. ولدراسة تركيب جينوم ما يتم أولا عمل خرائط جينية له ومنها تعمل خرائط جزيئية ثم يتم كولنة كل قطعة متداخلة (contigs) والتى فى مجموعها تغطى جزء من الجينوم ويتبع ذلك فك تتابعات كل منها - اي يجب هدم الجينوم إلى وحدات أصغر فأصغر حتى نصل إلى قطع تسمح للأساليب المعملية بفك تتابعاتها فيما يسمى [Top-Down approach for molecular mapping] وشكل (٢-١) يوضح تلك الخطوات كما اقترحها Klug & Cummings سنة ١٩٩٩. وإذا ما وصلنا إلى هذه النقطة، وجب علينا أن نعكس العمل وأن نعيد البناء مستعملين تلك القطع الصغيرة لنصل إلى الأكبر فالأكبر حتى الجينوم فيما يعرف بأسم [Bottom-Up approach].



شكل (٢-١) : رسم يوضح خطوات دراسة جينوم ما.

٢.٢. الخرائط الجينومية Genomic maps.

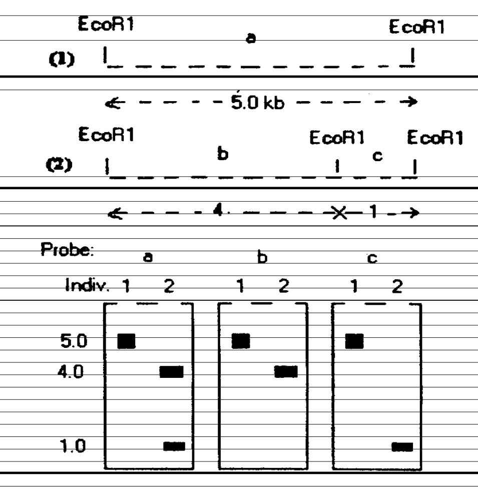
عرف علماء الوراثة الخرائط الجينية منذ بداية العشرينات من القرن الماضى، وكانت تعتمد على نسب العبور بين المواقع الجينية (loci) المختلفة التى تتعدد عليها الاختلافات المظهرية (phenotypic variants). هذه الاختلافات (الطفرات) عند جمعها بكائن ما، فإنها تكون عادة مقللة للحيوية أو حتى مميتة، كما أن هذه الطريقة تحتاج لعدد كبير من التهجينات - أو دراسة عدد كبير من سجلات العائلات الإنسانية. هذه الأسباب قللت من إمكانية استخدامها على نطاق واسع مع العلم بأن قدرة هذه الخرائط لا تسمح إلا بتحديد مواقع على الخريطة تبعد عن بعضها البعض بأكثر من 1 Mb. ولا يتوفر الآن سوى عدد قليل من البرامج التى تساعد فى عمل هذا النوع من الخرائط مثل برنامج MAPMAKER المتوفر بواسطة معهد Whitehead Inst., MIT بالولايات المتحدة الأمريكية. وظل الحال على ما هو عليه دون تقدم ملحوظ إلى أن اكتشفت الوسومات الجزيئية (molecular markers)، حيث تغير الحال وأمدتنا بعدد غير متناهى من الاختلافات على مستوى تتابعات الـ DNA حيث تتميز بتعدد الأشكال المظهرية (polymorphic) كما أنها متعادلة (neutral) التأثير على حيوية الفرد وتنعزل تبعا

للقواعد المندلية. استعمال تلك المسومات دفع بدراسة الخرائط الجينومية دفعة كبيرة إلى الأمام مكنتنا من زيادة قوة الإيضاح إلى حوالي 100 kb ولكن مثل تلك الأحجام من الجينوم مازالت كبيرة وغير مناسبة لفك تتابعاتها، لذلك وجب اللجوء إلى أساليب إضافية للمساعدة فى هذا المجال.

٢.٢.١. عمل الخرائط ببصمة إنزيمات القصر. Restriction fingerprint mapping

عادة تقطع اجزاء الـ DNA إلى شظايا صغيرة بواسطة عدد من إنزيما ت القصر لتكون مناسبة لربطها بأدوات النقل (cloning vectors) مثل البكتريوفاج لامدا الذى يحمل حوالى (20 kb) أو الكوزميدات تحمل حوالى (40 kb) أو الـ YAC الذى يحمل حوالى (200 kb)، ومع استخدام المجسات الجينية (probes) للتعرف على تلك الشظايا. فعلى سبيل المثال لو استعملنا انزيم القصر EcoR1 لقطع قطعة كبيرة من الـ DNA فأن احجام تلك القطع سيتوقف على اماكن تعارف هذا الأنزيم كما هو مبين بشكل (٢-٢)، ففى فرد ما (رقم ١) ستنتج قطعة قدرها 5 kb لوجود موقعين للقطع بينما وفى فرد آخر (رقم ٢) ستنتج قطعتين قدر الأولى 4 kb والثانية 1 kb نتيجة لوجود ثلاث مواقع لقطع الأنزيم فى ذات الموقع (تعدد اشكال ظاهرية). لعزل تلك القطع والتعرف عليها يمكن استخدام المجسات الجينية (probes) ولتكن فى هذه الحالة المجسات a و b و c التى تمثل الشظايا المتوقع الحصول عليها بعد تعليمها إما بالنظائر المشعة أو بمركب فلوروسنتى. وعند الفصل الكهربى للـ DNA على الأجاروز وبالتهجين مع المجسات تبعا لطريقة Southern blotting فستظهر الحزم fragments المبينة بشكل (٢-٢) حسبا للفرد والمجس المستعمل.

هذه الشظايا العديدة التى تم كولنتها وفك تتابعاتها تغطى جزء من جينوم ما ولكنها متداخلة مع بعضها البعض (overlapping) لذلك تعرف بشظايا الـ DNA المتداخلة أو ما يسمى بـ contigs. وإذا نجحنا فى تحديد أماكن بعض منها على الجينوم فإنها تصبح نوعا خاصا من المجسات الجينية تعرف بالتتابعات المحددة لموقع بعينه (sequence tagged sites) وتعرف أختصارا بأسم STS.



شكل (٢-٢) : رسم يوضح خطوات قطع الـ DNA بأنزيم القصر EcoR1 مع تمييزه باستخدام المجسات probes لجين ما.

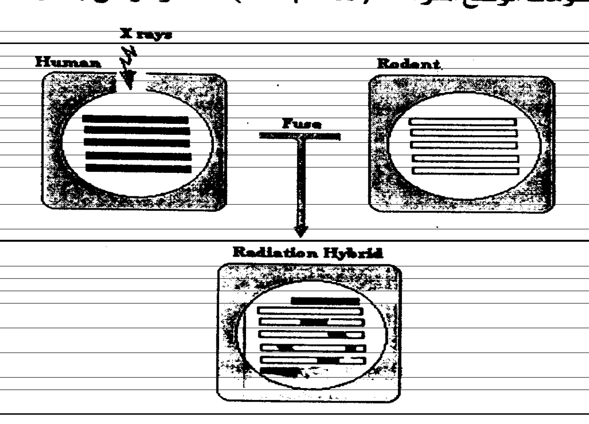
٢. ٢. ٢. طريقة دمج الخلايا بعد الإشعاع Radiation Hybrid (RH)

استغل العلماء النجاح فى التهجين أو الدمج ما بين الخلايا الجسمية (somatic cell hybridization) واستعملوها لتحديد أماكن الجينات على الكروموسومات. وقد استعملت خلايا لإنسان بعد إشعاعها بجرعة من أشعة x قدرها ٣٠٠٠ راد لتكسير الكروموسومات إلى شظايا صغيرة، ثم تدمج (fuse) تلك الخلايا المشعة مع خلايا الفأر، كما هو مبين بشكل (٢-٣). وبفحص الخلايا المهجنة (المدمجة) يلاحظ ارتباط شظايا إنسانية بكروموسومات الفأر بالإضافة لقطع غير مندمجة، وبحساب نسب تكرار الأندماجات المفردة والمزدوجة والتى تمثل الشظايا المتجاورة يمكن تقدير المسافات بينها، حيث وجد أن وحدة centi Rays أو (cR_{3000}) تساوى 0.1 cM لهذا النظام. وقد أقترح Griffith ومساعدوه سنة ١٩٩٩ أن إستعمال من ١٠٠ إلى ٢٠٠ خلية مدمجة من هذا النوع ستكون كافية لعمل خرائط قدرة إيضاحها عشرة أضعاف الخرائط الوراثية التقليدية.

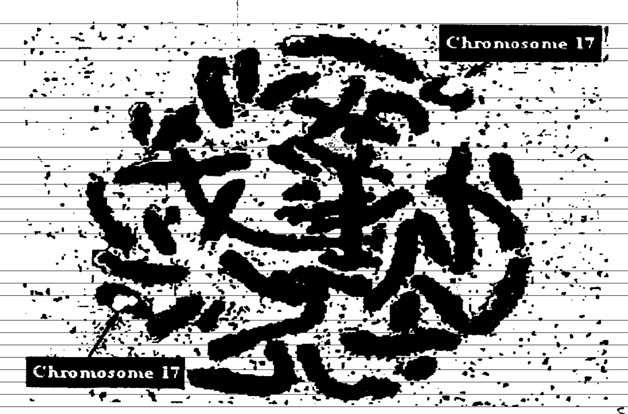
٢ . ٢ . ٣. طريقة استعمال المجسات على المستوى الخلوى.

Fluorescence *In situ* Hybridization (FISH).

ان استخدام طريقة التهجين على المستوى الخلوى بالمجسات الجينية المعلمة بمركبات فلوروسنتية، وتعرف اختصارا بأسم FISH، يمكن استخدامها أيضا لتحديد بعض المواقع على كروموسومات الوضع المتوسط (metaphase)، كما هو موضح بشكل (٢-٤).



شكل (٢-٣) : رسم توضيحى لطريقة RH فى الإنسان.



شكل (٢-٤) : صورة لكروموسومات الوضع المتوسط للإنسان توضح ارتباط مجس محدد على موضع بكروموسوم ١٧ باستخدام طريقة FISH.

والآن يمكن استخدام أكثر من مجس واحد للتحضير الواحد وذلك عند استخدام مجسات معلمة بمركبات فلوروسنتية ذات ألوان مختلفة بما يسمى chromosome painting. وقد اتضح أن استعمال كروموسومات الوضع المتوسط يؤدى إلى خرائط ضعيفة الوضوح ويرجع ذلك إلى مستوى التكاثف (condensation) والالتفاف الذى يتعرض له الـ DNA خلال هذا الدور من الانقسام، مما لا يتيح تحديد للمواقع التى يفصلها عن بعضها البعض مسافات تقل عن عدة ملايين من القواعد. ولزيادة كفاءة هذه الطريقة اقترح استخدام خلايا الدور البيني (intraphase)، كما طورت الطريقة بعد ذلك عن طريق فرد خيوط الـ DNA كيمائيا فى التحضيرات السيتولوجية قبل تثبيتها وصبغها فيما يسمى بطريقة (fiber-FISH) طبقا للعالم Hliskanen و مساعدوه سنة ١٩٩٥ وبذلك أمكنهم تحديد مواقع لا يبعدها عن بعض سوى 100 - 500 kb.

٢.٣. المكتبات الجينومية. Genomic libraries

اتفق العلماء على أن مصطلح المكتبات الجينومية، يقصد به مجموعة الكلونات (clones) التى تمثل جزءا اعتباطيا من الجينوم. وهذا العمل يحتاج إلى مجهود مضني مستخدمين العديد من التقنيات الحديثة، وفيما يلي الخطوط العريضة التى يجب إتباعها لعمل مكتبة لجينوم خاص بكائن ما.

٢.٣.١. تحديد عدد الكلونات المطلوبة.

Determining the number of clones needed.

ان تحديد عدد الشظايا (fragments) من الـ DNA التى يجب كولنتها يعتبر من المهام التمهيدية الهامةالتى يجب الأهتمام بهاقبل البدء فى عمل المكتبة. فأن استخدمنا عدد قليل من الكلونات فإنه بالتبعية لن يغطى القطعة الجينومية بالكامل، وإن استخدمنا عدد كبيرا من الكلونات فإنها ستغطى مناطق من القطعة الجينومية عدة مرات (تكرار) مما يعتبر إهدارا للجهد والمال. والمعادلة التالية توفر لنا حساب عدد الكلونات المطلوبة.

$$N = \frac{\ln(1-P)}{\ln(1-f)}$$

حيث

N = عدد الكلونات المطلوبة للحصول على كلون واحد لكل جين على هذه القطعة الجينومية عند مستوى احتمال P قدره ٩٩ ٪.

f = النسبة لمتوسط حجم الشظية بالكلون إلى متوسط حجم الجينوم الكلى.

فعلى سبيل المثال: أنا كان متوسط الشظية التى يحملها الفاج لامدا حوالي 0.02 Mb، ومتوسط حجم الجينوم الكلى لنبات الأرابيدوبسس حوالي 70 Mb فإن:

f = 2.86 x 10-4. & N = 1.61 x 104.

ويمكن تبسيط تلك القيم، بالقول بأن تقسيم جينوم الأرابيدوبسس إلى شظايا قدر كل منها 20kb بحيث يمثل كل جين بكلونة واحدة سنحتاج إلى ٣٤٩٧ كلونة فى حالتنا هذه. أي لتحديد كل جين على هذا الجينوم عند مستوى احتمال قدره ٩٩ ٪ فعلينا فحص عدد $١,٦١ \times ١٠^{٤}$ كلونة

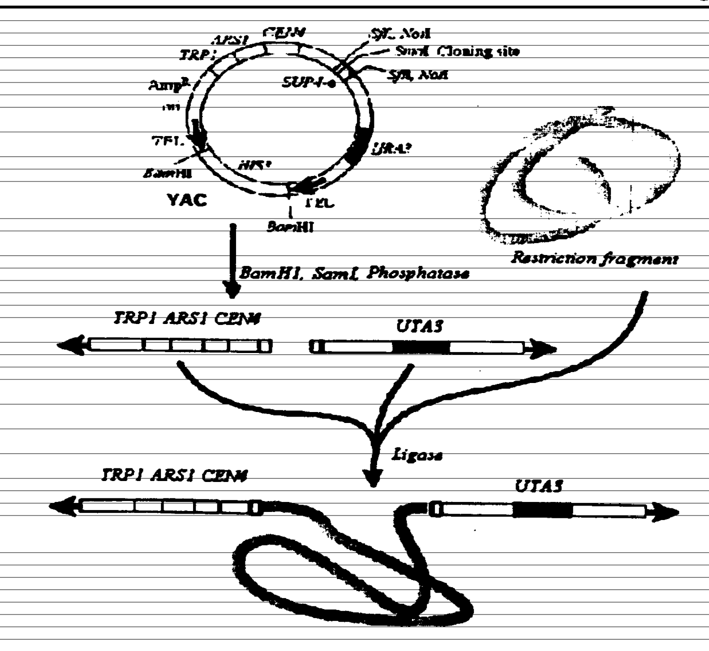
٢.٢.٢. أدوات الكولنة. Cloning Vehicles

تقدم لنا تقنيات الوراثة الجزيئية العديد من الأدوات والأساليب التى تساعدنا فى كولنة الشظايا الجينومية داخل أدوات للنقل لعمل المكتبات الجينومية، و لمزيد من التفاصيل فى هذا الصدد يمكن الرجوع إلى (ابويوسف و المتينى - ٢٠٠٢). وبطبيعة الحال فكلما زاد حجم القطع المكولنة كلما قلت الحاجة لأعداد أكبر من الكلونات لتغطية الجينوم تحت الدراسة. ويعتبر كروموسوم الخميرة الصناعى (yeast artificial chromosome) والمعروف اختصارا بأسم YAC من انسب الأدوات لهذا الغرض لقدرته على حمل شظية كبيرة من ١٠٠ - ١٠٠٠ قاعدة (فى المتوسط ٢٠٠ قاعدة)، كذلك فإن YAC يمكن أن يتضاعف كبلازميد فى البكتيريا *E. coli* و ككروموسوم فى الخميرة. ويتميز بتتابعات خاصة مثل منطقة السنترومير [*CNE*] و منطقة التضاعف الذاتى [*ARS*] و منطقة التلومير التى يمكن فتحها بأنزيم BamH1 [TEL] و العديد من المواقع التى تساعد فى إنتخاب الكلونات مثل [TRP ، URA3 ، His3] بالأضافة

لمواقع التعايش فى البكتيريا [Amp , ori]. شكل (٢-٥) يمثل رسما توضيحيا لطريقة ربط قطعة جينومية مع الـ YAC.

٢.٣.٣. ترتيب المكتبات. Ordering of Libraries

المكتبة الجينومية التى حصلنا عليها حتى الآن عبارة عن كلونات موزعة اعتباطيا (غير مرتبة) لجينوم ما، لذلك يجب علينا الآن العمل على ترتيب تلك الكلونات بالبحث أساسا عن مناطق التداخل بينها أو بما يسمى بالـ contigs - أي الشظايا المكولنة بها مناطق متماثلة مع تمثيل المناطق غير المتداخلة بينها مرة واحدة فقط. وبترتيب المكتبة نكون قد خطونا خطوة كبيرة إلى الأمام، فمنها يمكن أن نحدد أماكن الجينات المرغوبة وعدد نسخ كل منها بل يمكننا أن نعرف الكثير عى التعقيد التركيبى للجينوم ككل.

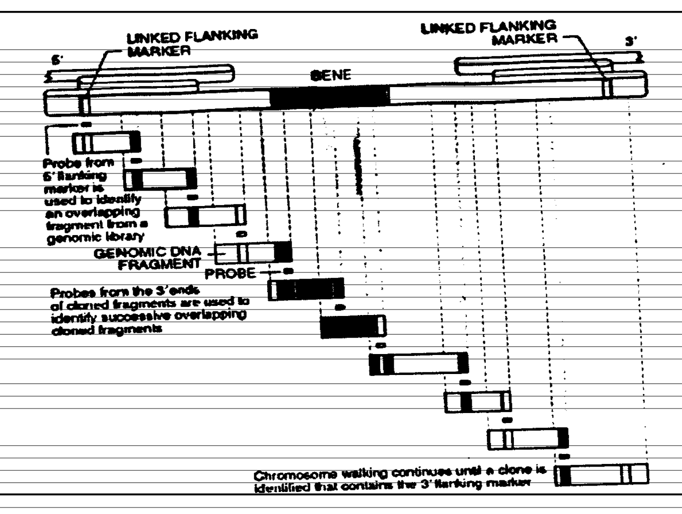


شكل (٢-٥) : رسم توضيحى لأسلوب ربط شظية جينومية مع الـ YAC كأداة للكولنة.

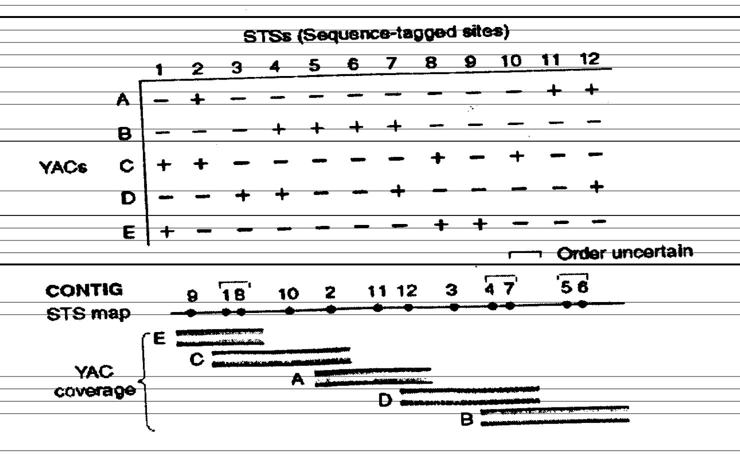
وأبسط الطرق لتحقيق ذلك يتم بواسطة البدء بكلونة ما من المكتبة تحت الدراسة، ثم البحث عن ثانية بها تتابعات متماثلة مع بعض من القطعة الأولى، ثم البحث عن قطعة ثالثة لها تتابعات متماثلة مع بعض من تلك بالقطعة الثانية ... وهكذا. وأول الطرق التى استغلت لتنفيذ ذلك وتعرف بأسم المشي الكروموسومى (Chromosome walking) أو القفز الكروموسومى (Chromosome jumping) وخطوات المشي الكروموسومى موضحة بشكل (٢-٦). يتبع الآن طرق عديدة أخرى خلافا للمشي الكروموسومى لترتيب المكتبات تعتمد على التهجين (hybridization) بين الجزيئات وعمل البصمات (fingerprints) للـ DNA عن طريق هدمه إلى قطع صغيرة، ولابد من استعمال برامج الكمبيوتر لأجراء آلاف المقارنات ما بين الكلونات لتحديد مناطق التداخل وترتيب الـ contigs. ومن الطرق شائعة الاستخدام طريقة بطاقات التتابعات المحددة الموقع على الجينوم sequence tagged sites أو ما يعرف اختصارا STS. وهى تتابعات صغيرة جدا أمكن تحديد مكانها على الجينوم، فمثلا إذا كانت الشظية (أ) عرف عليه الـ بطاقتين ١ و ٢ والشظية (ب) عرف عليها البطاقتين ٢ و ٣، بالتالي فإن الشظيتين أ و ب متداخلتين عند البطاقة ٢ وهكذا. وشكل (٢-٧) يوضح تلك الخطوات تبعا لـ Griffith ومساعدوه سنة ١٩٩٩.

٢. ٤ فك (سلسلة) التتابع Sequencing.

تقنية فك (أو معرفة) التتابع هى تلك الأساليب المعملية المتبعة لتحديد تتابعات النيوكلوتيدات الأربع على طول شظية صغيرة من الـ DNA. والأساس الذى يبنى عليه ذلك هى الطريقة الكيماوية الأنزيمية التى أقترحها Sanger منذ أكثر من ثلاثين عاما والتى ما تزال تستعمل حتى الآن وإن أدخلت عليها تعديلات لزيادة الكفاءة والاعتماد على التنفيذ الأوتوماتيكى. وتستعمل فيها النيوكلوتيدات الأربع معلمة بالنظائر المشعة أو الزوائد الفلوروسينتية مع النيوكلوتيدات المعدلة (dideoxynucleotides-ddN) التى توقف عمل أنزيم البوليميراز عند أرتباطها وبالتالى تحدد النيوكلوتيدة الأخيرة وهكذا عن طريق فصل الشظايا تبعا للحجم بالتفريد الكهربى، كما هو مبين بالشكل شكل(٢-٨). كما تتوفر الآن أجهزة إلكترونية مبرمجة يمكنها فك وتحليل التتابعات مباشرة و تقديم النتائج كعلاقات بيانية، كما هو مبين بشكل (٢-٩).

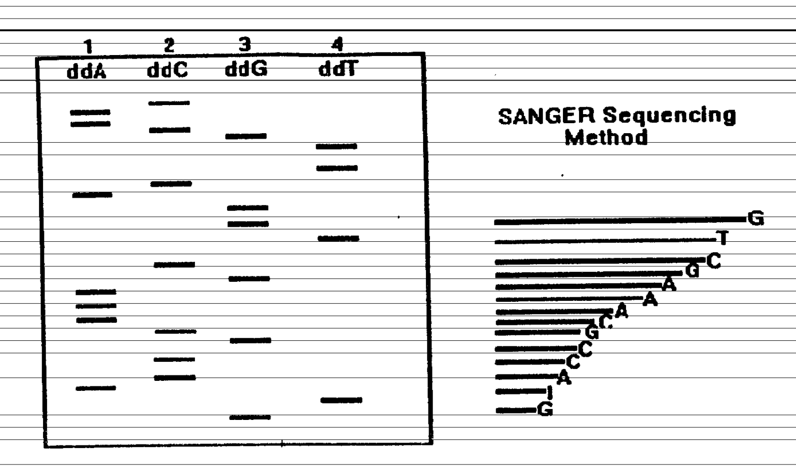


شكل (٢-٦) : خطوات عملية المشي الكروموسومي.

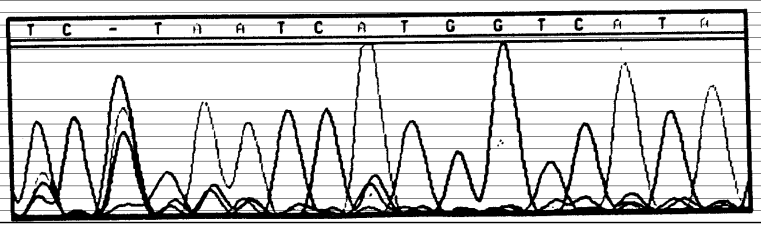


YACs	1	2	3	4	5	6	7	8	9	10	11	12
A	−	+	−	−	−	−	−	−	−	−	+	+
B	−	−	−	+	+	+	+	−	−	−	−	−
C	+	+	−	−	−	−	−	+	−	+	−	−
D	−	−	+	+	−	−	+	−	−	−	−	+
E	+	−	−	−	−	−	−	+	+	−	−	−

شكل (٢-٧) : رسم يوضح طريقة استعمال الـ STS لتحديد ترتيب الـ contigs .



شكل (٢-٨) : رسم توضيحى لطريقة سانجر لفك (سلسلة) التتابعات.



شكل (٢-٩) : نتائج سلسلة تتابعات عينة من الـ DNA بطرق أتوماتيكية.

عادة لسلسلة تتابعات شظايا جينومية ما نلجأ إلى أستراتيجيتين أساسيتين، الأولى تعتمد على فك تتابع القطع المكولنة المرتبة (ordered clone sequencing) ، والثانية هى فك تتابع الجينوم ككل (whole genome shotgun sequencing) والطريقة الثانية متبعة على نطاق واسع مع الجينومات الصغيرة للكائنات غير حقيقية النواة.

٢.٤.١. Ordered Clone Sequencing

ويمكن تلخيص هذه الأستراتيجية فى الخطوات التالية:

١. عمل مكتبة جينومبة مستغلين أداة نقل تتحمل شظايا كبيرة نسبيا.

٢. عمل خريطة جزيئية مرتبة لهذه المكتبة.

٣. عمل ما يسمى بخط السير الأدنى (minimum tilling path) ويقصد به العدد الأدنى الممكن من الـ contigs التى تغطى الجينوم بأكمله.

٤. أعادة كولنة هذه الـ contigs إلى شظايا أصغر حجما (2 kb) مستعملين الأداة المناسبة.

٥. فك تتابع الشظايا الصغيرة بأستعمال مناطق الأرتباط مع الأداة كبادئات (primers) ويفك التتابع فى كلا الأتجاهين بدئا من تلك المواقع.

٦. أعادة تجميع الشظايا (معروفة التتابعات) فى contigs مرة أخرى.

٢.٤.٢. Whole Genome Shotgun Sequencing .

ويمكن تلخيص هذه الاستراتيجية فى الخطوات التالية:

- عمل مجموعة عشوائية من الشظايا الصغيرة لجينوم ما عن طريق الهدم الميكانيكى (sonication) أو الهدم الأنزيمى ولكن تفضل الطريقة الأولى لتجنب التداخل الأنزيمى.
- التعرف على تتابعات أكبر عدد ممكن من تلك الشظايا الجينومبة، يمكن الاكتفاء بفك تتابع أطراف الشظايا فقط فيما يسمى (long insert library).

- تجمع تلك الشظايا (معروفة التتابع) فى contigs متداخلة، ولكن عادة ستتبقى فجوات (gaps) فيما بينها.
- ترتب contigs بالنسبة لبعضها البعض.
- تغلق الفجوات المتبقية تبعا لطريقة primer walking.

٥.٢. بطاقات التتابعات المميزة للجينات الفعالة.

.Expressed Sequence Tags [ESTs]

بطاقات التتابعات الفعالة والتى تعرف إختصارا بـ ESTs هى تتابعات من الـ DNA المبنية من جزيئات الـ DNA المكمل (cDNA) لرسالات الجينات التى عبرت عن فعلها بالنسخ فى نسيج ما تحت ظروف محددة. وكان العالم البيولوجى الشهير Venter قد اقترح فى أواخر الثمانينات من القرن الماضى على المعهد القومى الأمريكى للصحة (NIH) مشروعا بحثيا يدعوا إلى التركيز على التعرف على الجينات النشطة فقط فى الجينوم الأنسانى عن طريق التعرف على تتابعات مميزة للجينات النشطة فقط والمستخلصة من مرسالات (mRNA) الجينات التى نسخت بالأنسجة المختلفة، فيما أطلق عليه أسم ESTs، وبرر ذلك بأن الجينات الفعالة لا تمثل أكثر من ٣٪ من جملة الجينوم الأنسانى مما يجعل البحث عن جين ما مهمة صعبة للغاية، وأن تحديد موقع الجين فى حد ذاته لا يعكس طبيعة نشاط هذا الجين، كذلك فهو توفير للمال والجهد. ولكن رفضت الهيئة الأمريكية تمويل المشروع فى حينه مما دفع Venter إلى اللجوء إلى مجموعة من رجال الأعمال واقنعهم بتمويل إنشاء هيئة خاصة هدفها الأساسى التعرف على تلك البطاقات بالجينوم الإنساني ومحاولة استغلالها فى المجال الطبي، وتبعا لذلك أنشأ معهد أبحاث الجينوم (The Institute for Genome Research) والمعروف أختصارا TIGR [http://www.tigr.org] والذى أصبح بعد ذلك من أشهر الهيئات فى هذا المجال. ومنذ ذلك الوقت انتجت الـ ESTs على نطاق كبير، كانت تمثل الجينوم الإنساني فى أول الأمر ثم ما لبس أن تبعه العديد من الكائنات الأخرى، فحتى أكتوبر من سنة ١٩٩٩ تم تسجيل فى بنك الجينات (GenBank - http://www.ncbi.nlm.gov) حوالى ثلاثة ملايين بطاقة لعدد من الكائنات. وفى سنة ١٩٩٥ قام العالم Adams مع أكثر من تسعين باحثا فى نشر قائمة بهذه البطاقات فى الأنسان فى مجلة Nature زائعة الصيت حيث مثلت أكثر من ٨٠

الف جين إنسانى ٨٠٪ منها لم تكن معروفة من قبل، والجدول التالى يلخص هذه البطاقات لكل عضو درس حتى تاريخه.

مصدر الـ DNA	عدد البطاقات	مصدر الـ DNA	عدد البطاقات
الجلد	٣٠٤٣	خلايا الدم	٣٥٠٥
العظم	٥٧٣٨	الغدة الدرقية	٢٣٧٨
الغدة جارة الدرقية	١٩٧	الأغشية المصلية	٥٧٣٦
الكبد	٣٧٨٠٧	المرارة	٣٧٥٤
القولون	٤٨٣٢	العضلات الملساء	٢٩٧
الأمعاء الدقيقة	١٠٠٩	البروستاتا	٧٩٧١
الخصية	٧١١٧	المبيض	٣٨٤٨
الرحم	٦٣٩٢	المشيمة	١٢١٤٨
المخ	٦٧٦٧٩	العين	١٩٣٢
الغدد اللعابية	١٨٨	المرئ	١٩٤
الأنسجة الدهنية	٢٤١٣	القلب	٩٤٠٠
الغشاء البروتونى	٨٨٤	الطحال	٧٩٢٤
الكظر	٣٤٢٧	البنكرياس	٥٥٣٤
الكلي	٣٢١٣	البربخ	١٧١٦
العضلات	٤٦٩٣	الأغشية الزلالية	٣٨٨٩
الجنين	١٩٢٩١	الغدة الثيموسية	٢٤١٢

لتوضيح أهمية هذه البطاقات ESTs فى تحديد الجينات، دعنا ندرس مثال على ذلك، فبفرض أننا حددنا موقع جين ما بين الموقعين ٨٧ و ٩٣ وحدة عبورية سنتيمورجان (cM) على الكروموسوم رقم ١٦ فى الأنسان، فيمكننا الرجوع إلى الجداول الخاصة حيث يمكن أختيار بعض البطاقات ESTs التى تتواجد ما بين هذين الموقعين وفى أي الأعضاء هى نشطة، فوجد أن البطاقة رقم A006F26 ترتبط بهذا الموقع وهى

فى الأصل تظهر تماثلا كبيرا مع الجين الخاص بالحالة المرضية المعروفة بأسم
. Huntington's Chorea

٣- الجينومكس الوظيفى
Functional Genomics

إعداد : ياسر مبروك و أحمد المتينى

مع بدأ العمل فى مشاريع فك وتحليل الجينومات للعديد من الكائنات الأولية والراقية، ظهرت الحاجة إلى إبتكار وتوفير الأساليب المعملية وكذلك الأسس النظرية لمعرفة وتحديد وظائف وطبيعة عمل الجينوم ككل، وعليه ظهر هذا الفرع الجديد من العلوم البيولوجية الجزئية والذى يعرف باسم الجينومكس الوظيفى Functional genomics. ومما سبق يتضح لنا أن هذا الفرع يهتم بدراسة وتحديد أماكن النشاط الجينى فى الجينوم، ومن خلال هذه الدراسة يتوقع منا محاولة الإجابة على عدة تساؤلات :

(١) أين ومتى يتم التعبير عن جين ما ؟

(٢) ماهى طبيعة ناتج الفعل الجينى ؟

(٣) ما هو مدى التعاون بين الجينات لإظهار فعلها ؟

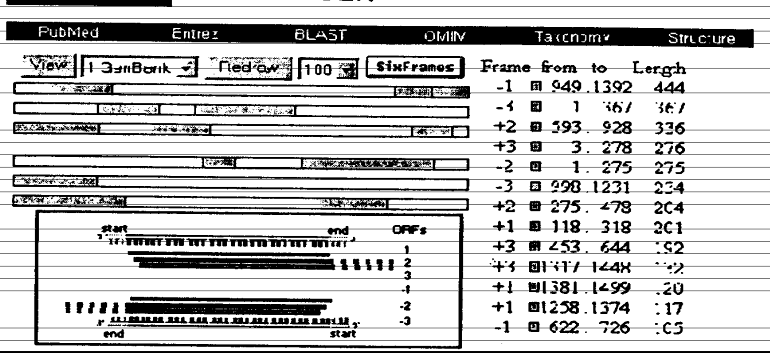
فى محاولة أجابتنا على السؤال الأول، لابد ان نتناول بالتدقيق مواضيع محددة مثل ما هى مواصفات المواقع التى تنسخ؟ وما دور عوامل التنظيمية وتداخلاتها على المناطق غير المنسوخة؟ هذا المجال من الدراسات يمكن أجماله فيما يسمى بالـ Transcriptomics – بمعنى آخر هو دراسة المرسالات mRNAs. أما التساؤل الثانى فسيختص بدراسة عمليات الترجمة وتكوين الناتج الجينى، ألا وهو البروتينات. هذا المجال من الدراسات يمكن إجماله بما يسمى بالـ Proteomics – وسوف نتناوله بشيء من التفصيل فيما بعد. أما تساؤلنا الثالث فيهتم بدراسة التداخل والتعاون بين الجينات لأظهار الشكل الظاهرى فيما يسمي بـ Phenomics – وقد نتناول هذا الموضوع بالدراسة من خلال مواضيع الوراثة التكوينية Developmental genetics.

١.٢. التنبؤ بمواقع الجينات .Predection of Gene Locations

إن الإطارات المفتوحة للترجمة (open-reading frames – ORFs) هو تعبير مرادف لمصطلح الجينات النشطة التى تنسخ وتترجم. وعادة يكون عدد الـ ORFs فى جينوما ما محدودة للغاية مقارنتا بباقى المادة الوراثية بالجينوم، ومع إزدياد عدد الجينومات التى مسحت وفحصت كان من الضروري محاولة البحث أو التنبؤ بهذه المواقع فيما يسمى ORFs prediction والتى لم تكن معروفة من قبل. وعادة لعمل هذا البحث نلجأ لبرامج الكمبيوتر المختصة اعتمادا على خواص الـ ORFsأو ما يطلق عليه مسمي الـ signals، فعادة تبدا بكودون " بدء الترجمة" وهو غالبا AUG ولابد ان ينتهى بأى واحدة من الكودونات الثلاثة الخاصة بالإنهاء، ويجب أن يكون طويل نسبيا (على الأقل ١٠٠ كودون) حتى يترجم إلى بروتين ذو معنى. وعلى ذلك فطول الـ ORF الإفتراضى سيساوى ٣/٦٤ حيث ٦٤ هو عدد الكودونات الكلى و٣ هو عدد كودونات الإنهاء. وفى أغلب الأحيان تستعمل برامج الكمبيوتر المبنية على النماذج الإحصائية خصوصا نماذج ماركوف (سنتناولها ببعض التفصيل بالباب القادم) للتنبؤ وتحديد هذه الإطارات من ضمن الإطارات الستة المحتملة لكل ORF كما هو موضح بشكل (٣-١). والاعتماد على النماذج الإحصائية فى كثير من الأحيان يؤدى إلى استبعاد معطيات هامة : مثل الأكسونات الصغيرة والتى يقل فيها عدد الأحماض الأمينية عن ١٠٠، كذلك تشخيص إطارات يزيد طولها عن ١٠٠ ولكنها تندرج تحت نوعية الجينات التى تنسخ ولا تترجم...... وهكذا العديد من الحالات المتداخلة، لذلك وجب تأكيد تقديرات الكمبيوتر وحساباته بأجراء التجارب المعملية المبنية على تقنيات البيولوجيا الجزيئية.

مما سبق، فإن هذا المفهوم يكون الأتجاه العام الذى سنتبعه فى تناولنا لموضوعات الجينومكس والمعلوماتية الحيوية. فلدراسة الفعل الجينى تحت مسمي الجينومكس الوظيفي يمكن أن نسلك أحد مسارين أوكليهما معا للوصول للهدف المرجو، الأول من خلال تطبيقات علوم الكمبيوتر واعتمادا على قواعد البيانات المتخصصة من خلال البحث والمقارنة والاستدلال والتنبؤ فيما يسمى بالبحث عن القرائن homology، وهى ما سنتناوله فى الأبواب التالية – أو ما يعرف بالمسار المعلوماتي. أما المسار الثانى – ألا وهو

المسار التجريبي تبعا لتقنيات البيولوجيا الجزيئية، فهو يشمل عدة اتجاهات يمكن إجمالها فى بعض الأساليب التالية.

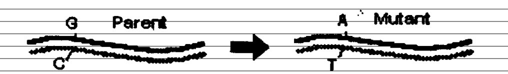


شكل (٣-١) : رسم توضيحي لأحد الجينات توضح الأطارات الستة المحتملة له بناء على أن الكودونات تتكون من ثلاث نيوكلوتيدات. بالأضافة لصفحة من برنامج ORF Finder المتوفر من قاعدة NCBI يوضح نتائج البحث عن هذه الأطارات لجين ما.

٣. ٢. أساليب الإطاحة بفعل الجين Gene Knockout.

هذا الأتجاه يعتبر من أكثر الأتجهات شيوعا لدراسة الفعل الجينومى، فعن طريق إسكات عمل جين ما يمكن تتبع التداعيات التى ستظهر على الشكل الظاهرى حيث سمى هذا الأتجاه بالوراثة العكسية reverse genetics لكونها عكس ما جرى عليه الحال فى الدراسات الوراثية التقليدية. ويرجع الفضل إلى العالم Michael Smith (نال جائزة نوبل للعلوم سنة ١٩٩٣) فى أواخر السبعينات من القرن الماضى لتطوير طريقة لأستحداث الطفرات داخل النظام الحي mutagenesis *in vivo* و المعروف أيضا باسم الطفور موجه لموقع site-directed mutagenesis لأستحداث طفرة موضعية عند موقع محدد من تتابع ما من الـ DNA قد يمثل ORF. وقد طورت تقنيات إستحداث الطفرات هذه فيما بعد واصبحت الطرق الحديثة أكثر كفاءة حتى وصلت أحتمالات النجاح الآن لأكثر من ٧٥٪ فى الحصول على الطفرة المرغوبة، فلو كان عندنا قطعة من الـ DNA تمثل ORF

ما (فدعنا نعرفها بالجزئ الأبوى parent molecule) و إذا ما كنا نرغب فى إدخال طفرة (إستبدال نيوكلوتيدات) على موقع محدد فيمكن تمثيلها كالتالى:



وهذا الأستبدال قد يكون عند الأطراف أو عند منتصف الجزئ.

٣.٢.١. إستبدالات substitutions عند اطراف الجزئ.

بأستعمال تقنيات تفاعلات البلمرة المتسلسلة polymerase chain reactions المعروف اختصارا بـ PCR يمكن احداث مثل هذه التغيرات. فمثلا لو هناك قطعة من الـ DNA تمثل تتابع ما كما هو مبين بالتالى:

5'TCTATGGACCAGTACGATACCAGTA.....CGACCTACGTAGACTAGACGGATAGAG 3'
3' AGATACCTGGTCATGCTATGGTCAT..... GCTGGATGCATCTGATCTGCCTATCTC 5'

وهذا هو التتابع الأصلى (الأبوى) parental molecule، فلتغير اطرافها يمكن أستعمال زوج من البادئات primers لتعظيمها amlification بواسطة الـ PCR، ويجب أن يكون تصميمها كالآتى:

البادئ من الشمال = '5' TCTATGGACCAGTACGAT 3

والبادئ من اليمين = '5' CTCTATCCGTCTAGTCTA 3

فبأستعمال هذين البادئين يمكن تعظيم تلك القطعة ملايين المرات، ولكن لو أريد أن تتضمن تلك القطعة تتابعات لبعض إنزيمات القصر مثل EcoRI (GAATTC) من الطرف الشمال ومثل BamHI (GGATCC) من الطرف اليمين فيمكن اضافة تلك التتابعات لكل من البادئين السابقين ليصبح البادئ الأيسر

5' **GCGAATTC**TCTATGGACCAGTACGAT 3'

والبادئ الأيمن 5' GCGGATCCCTCTATCCGTCTAGTCTA 3'

ويلاحظ اضافة تتابع GC من كلا الطرفين وذلك لتسهيل عمل القطع للأنزمين فيما بعد، وبأجراء الـ PCR مرة اخرى بأستعمال هذين البادئين فإن القطعة المعظمة amplified fragment ستصبح كالتالي:

GCGAATTCTCTATGGACCAGTA...
CGATACCAGTACGACCTACGTAGACTAGACGGATAGAGGGATCCGC
CGCTTAAGAGATACCTGGTCATGCTATGGTCAT.....
GCTGGATGCATCTGATCTGCCTATCTCCCTAGGCG

وبقطع هذه الجزيئات بأنزمي القصر EcoRI و BamHI سنحصل على القطع التالية:

AATTCTCTATGGACCAGTACGATACCAGTA....
CGACCTACGTAGACTAGACGGATAGAGG
GAGATACCTGGTCATGCTATGGTCAT.....
GCTGGATGCATCTGATCTGCCTATCTCCCTAG

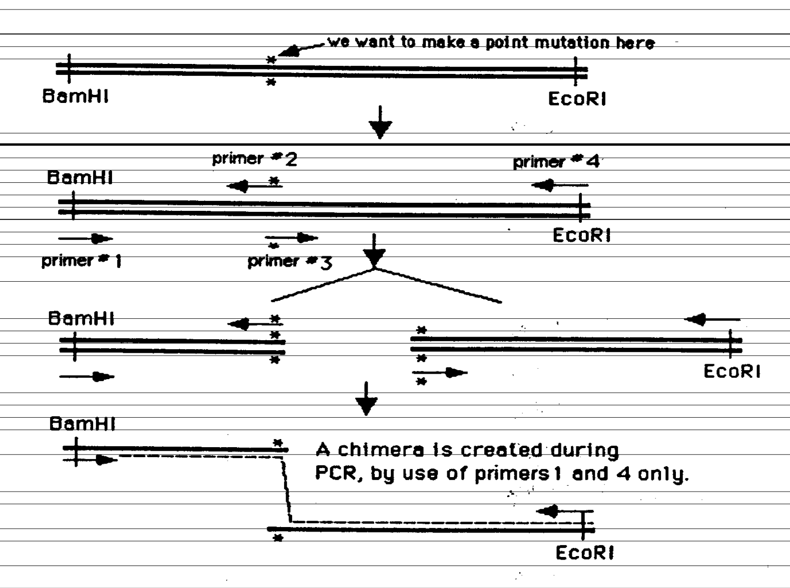
ويلاحظ أننا في النهاية قد حصلنا على قطعة جديدة من الـ DNA تم تحوير نهايتها عن طريق إدخال تتابعات من النيوكلوتيدات غير المتوافقة mismatch (طفرات إستبدالية) عن طريق تحوير البادئات.

۳. ۲. ۲. إستبدالات substitutions عند وسط الجزئ.

عادة الأستبدالات (الطفرات الموضعية) يرغب في تواجدها في وسط التتابع (لأسكات فعل ORF ما) ونادرا عند نهاياته، ومن اجل هذا نلجأ لإستعمال عدد أكبر من البادئات، كما هو موضح بشكل (۳-۲). فإذا ما رغبنا في إدخال طفرة بموقع متوسط والممثلة في الرسم بالعلامة "*" نستخدم ٤ بادئات مختلفة منها أثنين محيطة بالقطعة وهما primer#1 و primer#4 والأثنين الآخرين يغطيان منطقة التغيير (الطفور) وهما primer#2 و primer#3، وبتعظيم قطع الـ DNA هذه سيتولد لنا قدر كبير من القطع كل يمثل نصف القطعة الأولية parent molecule والتي تتميز أطرافها بمواقع لتعارف بعض إنزيمات القصر (لتسهيل عمليات الكلونة cloning فيما بعد). وفي الخطوة التالية تستعمل

القطع الجديدة فى دورة جديدة من التعظيم بواسطة الـ PCR ولكن فى هذه المرة سنستعمل البادئات primer#1 و primer#4 فقط وفى هذه الحالة يمكن أن يتولد لدينا قطع تمثل خليط من النصفين فيما يعرف بأسم chimera molecule حيث سيمثل الجزئ الأصلى بعد إدخال بعض الإستبدالات التى تؤدى إلى عدم توافق mismatch وذلك فى منتصفه (كما هو مبين بشكل: ٣-٢).

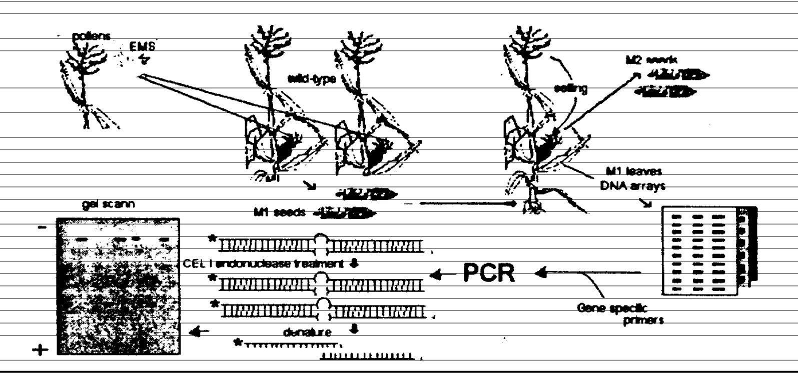
عادة "تكلون" مثل هذه القطع الجديدة (الطافرة) فى بلازميدات وتستغل فى تجارب للتهجين بغرض الحصول على أفراد عبورية recombinants بها التتابعات الطافرة وخلفية من التراكيب البرية wild-type.



شكل (٢-٢): رسم توضيحى يمثل خطوات إستحداث طفرة بمنتصف قطعة من الـ DNA.

٣. ٢. ٣. استعمال المطفرات الكيميائية مع تقنيات الـ PCR.

الطرق السابقة لإستحداث طفرات موجهة الموضع site-directed عادة يمكن تطبيقها بسهولة نسبية فى الكائنات ذات الجينومات الصغيرة أو المعروفة بالتحديد، أما الكائنات كبيرة الجينوم مثل نبات الذرة maize plants فكان تطبيق تلك الطرق السابقة بها غاية فى الصعوبة حتى أقترح Till وزملاءه سنة ٢٠٠٤ طريقة جديدة تجمع مع أستعمال المطفرات الكيميائية (مثل EMS) وتقنيات الـ PCR و تعرف هذه الطريقة بأسم TILLING وهو إختصار لـ Targeting Induced Local Lesions IN Genomes . وفى هذه الطريقة، تجمع حبوب اللقاح pollens من أحد نباتات الذرة التى تمثل صف line ما وتعامل بالمطفر القوى EMS (ethylmethansulfonate) ثم تلقح كيزان النباتات الطبيعية (والتى لها نفس الخلفية الوراثية للنباتات المعاملة) بحبوب اللقاح المعاملة، ثم زرع الحبوب نباتات الذرة التى تمثل صف line ما وتعامل بالمطفر القوى EMS (ethylmethansulfonate) ثم تلقح كيزان النباتات الطبيعية (والتى لها نفس الخلفية الوراثية للنباتات المعاملة) بحبوب اللقاح المعاملة، ثم تزرع الحبوب الناتجة للحصول على نباتات الجيل الأول الطافر M_1-generation. يمكن بالتلقيح الذاتى الحصول على نباتات الجيل الثانى الطافر M_2-generation وهكذا. نباتات الجيل الأول M_1 تعتبر خليطة وراثيا heterozygous لأى طفرة مستحدثة، ولتحديد الطفرات تجمع عينات من أوراق الجيل الأول او صفوف الجيل الثانى حيث يستخلص الـ DNA من أنسجة الورقة لكل نبات على حدة، ثم تستعمل بادئات primers خاصة بالجين المراد إدخال طفرات عليه، ثم تعظم بواسط الـ PCR. تجمع العينات المعظمة وتفصل الخيوط denatured عن بعضها ثم يعاد الأرتباط بينها مرة أخرى re-annealed لتكوين جزيئات خليطة من الخيوط طافرة وخيوط طبيعية heteroduplexes،ثم تعامل هذه الجزيئات الخليطة بأنزيم القصر CEL I، ثم تفحص القطع الناتجة بالتفريد الكهربى حيث يمكن تحديد الصفوف الطافرة من غيرها من حجم القطع المفرودة، كما هو موضح بشكل (٣-٣).

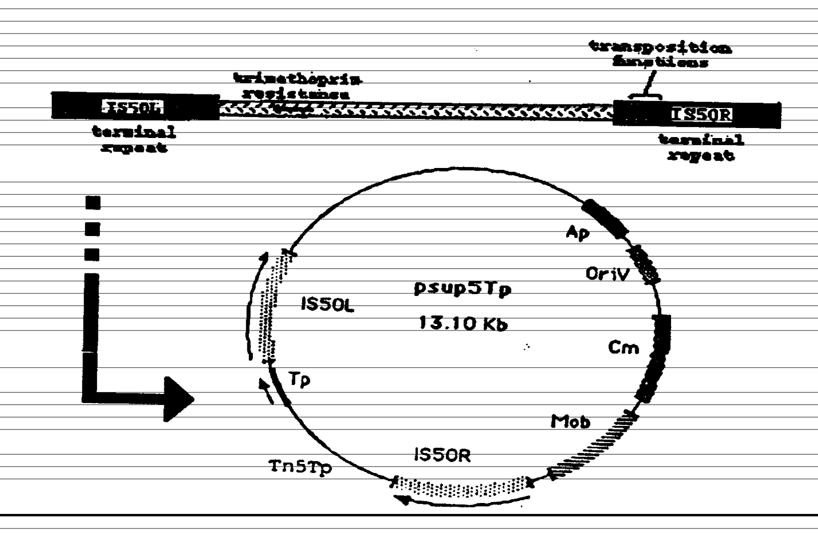


شكل (٢-٢) : رسم توضيحى لخطوات عزل الطفرات بنبات الذرة بأستعمال طريقة TILLING.

٣. ٢. ٤. التطفر بالإقحام Insertional mutagenesis .

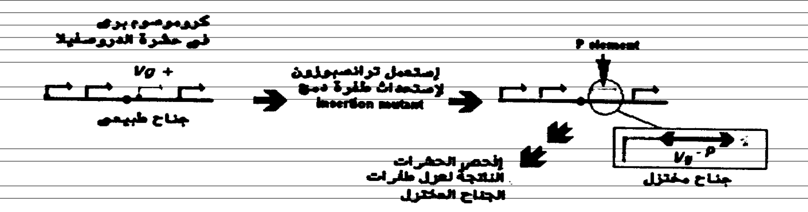
بالأضافة للطرق السابقة لإستحداث طفرات محددة الموقع تستعمل العوامل الوراثية المتحركة mobile genetic elements، خصوصا الترانسبوزونات Transposons، لقدرتها على الإندماج insertion مع DNA خلية العائل host cell الذى قد يكون كروموسوما أو فيروسا أو بلازميدا أو غيرها، وبإندماجها تؤدى إلى إضطراب الجين عند موضع الدمج مسببة فقده لقدرته على العمل فيما يندرج تحت مجموعة طفرات الإطاحة knockout mutants. وقد أتبعت هذه الطريقة فى بادئ الأمر فى البكتريا بنجاح حيث استعمل الترانسبوزون المعدل Tn 5Tp على نطاق واسع، حيث ادخل جين المقاومة للمضاد الحيوى trimethoprim محل الجين الخاص بمقاومة المضاد kanamycine فى الطرز البرية، كما هو مبين بالشكل (٣-٤). وعادة ما يولف هذا الترانسبوزون مع بلازميدات معروف بها عدد من جينات مقاومة المضادات الحيوية مثل الـ Amp = Ampicilline و الـ Cm = chloramphenicol، بالأضافة لجين مقاومة المضاد Tp = trimethoprim الموجود على الترنسبوزون Tn5Tp نفسه. وعادة يتم التزاوج بين البكتريا *E. coli* الحاملة لهذا البلازميد المركب مع البكتريا المراد إستحداث الطفرات بها ثم يفحص النسل الناتج لتحديد الخلايا المقاومة للمضاد Tp كدليل على إندماج

الترنسبوزون بها ثم تفحص هذه المجموعة للبحث عن الطفرات المستحدثة أو المرغوبة. ولكن من المصاعب التى واجهاتها هذه الطريقة كانت قدرة الترانسبوزون على التنقل بفعل إنزيم الـ transposase مما يؤدى لأرتداد الطفرات وفقدها بسهولة.



شكل (٢-٤) : رسم توضيحى يمثل الترانسبوزون Tn5Tp منفردا، وبعد ربطه بالبلازميد المركب pSUP5Tp (حيث الجينات، Amp - مقاومة الأمبسلين و Cm - مقاومة الكلوروامفنيكول و Tp - مقاومة تريمثوبريم و OriV - مبدا التضاعف و mob - مبدا النقل و الترانسبوزون Tn5Tp).

فيما بعد عرفت العديد من الترنسبوزونات فى الحيوانات خصوصا الدروسفيلا والفئران، وقد استعمل العامل الوراثى المتنقل المعروف بأسم P element على نطاق واسع لإستحداث طفرات بالنسيج الجنسى لحشرة الدروسفيلا بناء على الخطوات المبينة بشكل (٢-٥).

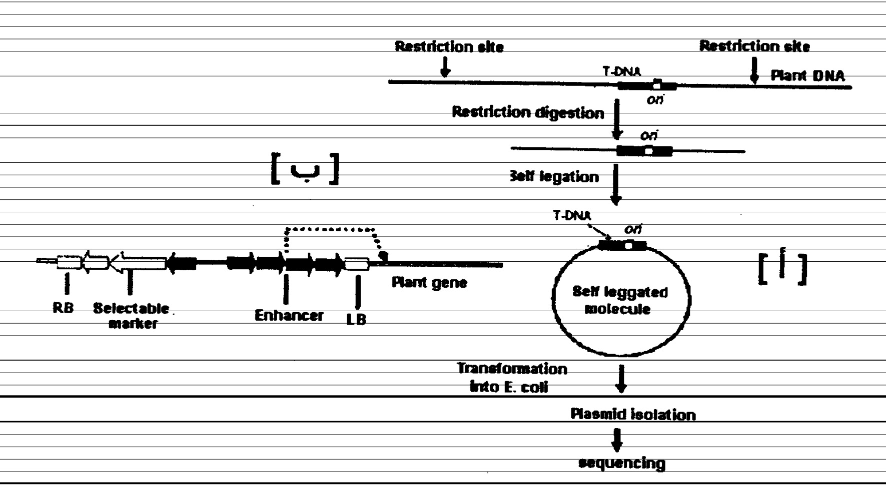


شكل (٢-٥) : رسم تخطيطى لإستحداث طفرات لصفات الجناح فى الدروسفيلا بأستخدام العامل الوراثى المتنقل P.

لكن كان يعيب هذه الطريقة هى الأخرى عدم الثبات بسبب تنقل الترنسبوزونات بالأضافة إلى أن العامل P فى الدروسفيلا يندمج بمواقع محددة على الكروموسوم مما يشكك فى عدم حيادية إستحداث مثل تلك الطفرات. ولقد أستخدم فى الدروسفيلا ترانسبوزونات أخرى لتجنب تخصصية مواقع الدمج مثل الترانسبوزونات المعروفة بأسم Hermes and piggyback.

اما فى النباتات الراقية خصوصا تلك التى تم الإنتهاء من التعرف على جينوماتها الكاملة مثل نبات الـ Arabidopsis ونبات الأرز فتستحدث بها طفرات الأقحام هذه بواسطة قدرة الأندماج لقطع الـ T-DNA المحورة والخاصة بالبكتيريا *Agrobacterium tumefacians* المستعملة على نطاق واسع فى تجارب التوليف الوراثى بالنباتات. و نبات الـ Arabidopsis هذا يعتبر نموذجا لمثل تلك التجارب، فقد استعمل على نطاق واسع لهذا الغرض لمعرفتنا المستفيضة عن خرائط تتابعاته والواسمات الجزيئية العديدة المتوفرة بجينومه، وقد أستعمل الـ Arabidopsis من خلال عدة تقنيات للتطفر تؤدى إلى فقد عمل الجين loose of function عن طريق الأقحام والتى تعرف بأسم التذييل بواسطة الـ T-DNA اى T-DNA tagging. أن إستعمال قدرة الأقتحام او الأندماج التلقائية للـ T-DNA فى تتابعات جينوم الـ Arabidopsis الصغير وبواسطة

الكلونة وتقنيات الـ PCR يمكننا معرفة التتابعات المحيطة بجانبي منطقة الأقحام، كذلك يمكننا التنبؤ بالجين الذى فقد نشاطه نتيجة للأقحام. ولكن يجب التأكيد على أن جملة الطفرات المظهرية المشاهدة فى مثل تلك التجارب لا يمكن ارجاعها جميعا لعملية الأقحام ذاتها لذلك وجب التوصل لطريقة لتأكيد الأرتباط بين الطفرة المشاهدة وعملية الأقحام. شكل (٣-٦) يوضح احد الأستراتيجيات المستعملة لذلك والتى تعرف بأسم plasmid rescue حيث يدخل تتابع منشأ التضاعف *ori* على تتابعات الـ T-DNA حتى يمكنه الدخول والتضاعف فى البكتريا *E. coli*.



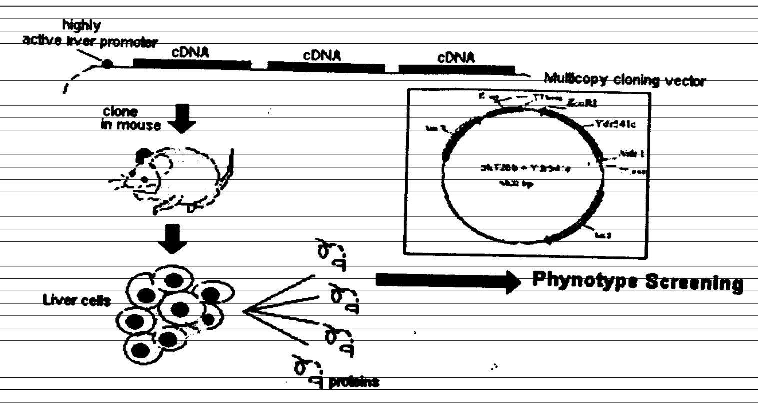
شكل (٣-٦) : [أ] رسم تخطيطى لخطوات طريقة الـ Plasmid rescue فى نبات الـ Arabidopsis. [ب] رسم توضيحى للـ T-DNA المدخل علية تتابعات المستحث enhancer لتصيد الجينات.

ونتيجة للجهود الكبيرة فى هذا المجال خلال السنوات القليلة الماضية، يمكننا الآن الحصول على عدد كبير من سلالات الـ Arabidopsis الطافرة بالأقحام insertion-mutated lines والتى يبلغ عددها حوالى ١٧٥٠٠٠ سلالة مختلفة، يمكن الحصول عليها من مراكز حفظ الأصول العالمية Arabidopsis stock centers كما فى جامعة أوهايو بالولايات المتحدة أو جامعة نوتنجهام ببريطانيا. وهناك تقنية مماثلة ولكنها تعتمد على إستحداث طفرات بالأقحام ولكن هذه الطفرات ترجع لظاهرة إسترجاع النشاط الجينى gain of function وليس فقده. وتعتمد هذه الطريقة المعروفة بأسم "مصيدة الجينات" gene trapping أعتمادا على تراكيب مولفة من الـ T-DNA مدخل بها تتابعات من المحثات للنسخ من بروموتور قوى مثل البروموتر CaMV 35S، كما هو مبين بشكل (٣-٦ب). فمثل هذا الأقحام قد يستحدث طفرات مظهرية جديدة لأحد أفراد أحد العائلات الجينية gene family والتى كانت غير عاملة.

٣.٢. الفعل الجينى الفائق Gene Overexpression.

هذه التقنية مثلها مثل سابقاتها صممت لدراسة وظائف الجينات وهى تعتبر معاكسة للمفاهيم التى تناولناها فى التقنيات السابقة للإطاحة بفعل الجين فهى على العكس تهتم بدراسة تأثير النشاط الجينى الزائد على الشكل المظهرى. وفى هذه الطريقة، يتم إدخال عدد كبير من نسخ الجين تحت الدراسة إلى بلازميد أو أداة من أدوات النقل الجينى multicopy cloning vector، كما هو موضح بشكل (٣-٧). يتم بعد ذلك إدخال هذه الجينات فى جينومات الكائنات تحت الدراسة، ومنها الفئران لدراسة النشاط الجينى الزائد على الشكل الظاهرى.

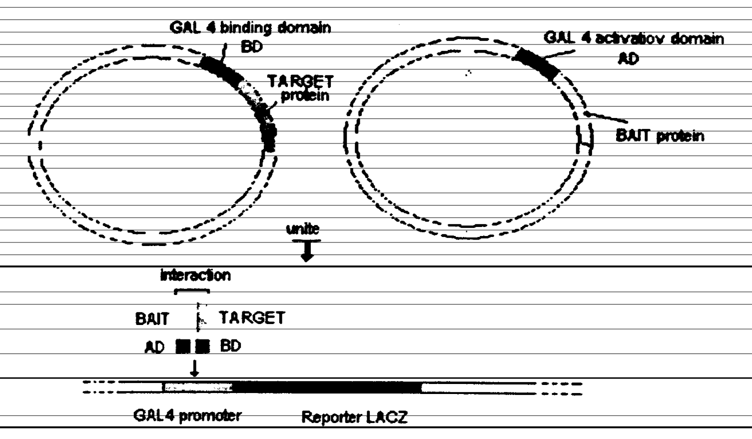
وتستعمل هذه التقنية كثيرا لدراسة العلاقة بين الجينات ومرض السرطان cancer فى الثدييات.



شكل (٢-٧) : رسم توضيحى لكلونة فأر بأستعمال أدات متعددة النقل الجينى.

٣. ٤. دراسة التداخلات بين البروتينات Proteins Interactions.

لدراسة التداخل interaction بين البروتينات لمعرفة مدى الأعتماد الوظيفى بينها يستعمل التكنيك المعروف بأسم yeast dihybrid حيث يستعمل بلازميدين، فعلى سبيل المثال الجين المنظم للنسخ المعروف بأسم Gal 4 يتكون من وحدتين بروتينتين two domains يشفر لكل منهما تتابعات محددة على الجين، أحدهما خاصة بالتنشيط activation والثانية بالأرتباط بتتابعات خاصة على الـ DNA ولابد من ارتباط الوحدتين مع بعضهما بالتوازى juxtaposition لكي يستطع القيام بعمله. ففى البلازميد الأول "يكلون" الجزء الأول للجين Gal 4 والمسئول عن تتابعات التنشيط ويربط معه تتابع الجين المسئول عن البروتين الأول المراد دراسته bait protein، أما البلازميد الثانى فيكلون به الجزء الثانى من الجين Gal 4 والمختص بالأرتباط بالـ DNA مع البروتين الثانى تحت الدراسة target protein، كما هو موضح بشكل (٢-٨). ويدمج البلازميدين مع بعضهما فإذا ما كان كلا البروتينين تحت الدراسة يمكنهما الأرتباط مع بعضهما، فإنهما بالتالى سيسمحان بالأرتباط بين وحدتى البروتين Gal 4 وبالتالى يكون قادر على تنشيط البروموتر لنسخ جين إستدلالى reporter gene وهو فى هذه الحالة Lac Z فإذا ما أظهر فعله دل ذلك على أن البروتينين تحت الدراسة يتداخلان مع بعضهما البعض.



شكل (٢-٨) : رسم توضيحى لطريقة الـ Yeast dihybrid method.

٥. التعارض للإسكات الجينى بواسطة جزيئات الـ RNA الصغيرة - RNAi

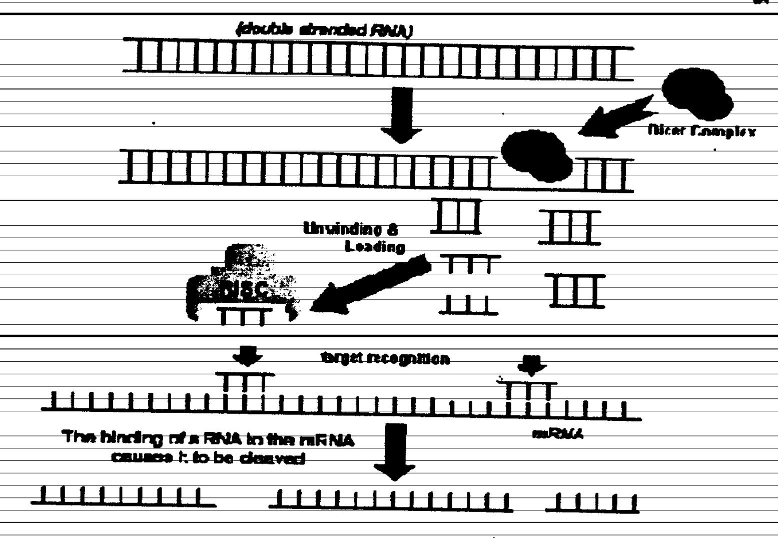
عرفت هذه الظاهر منذ وقت ليس ببعيد، عندما لوحظ فى منتصف الثمانيات من القرن الماضى، أن نباتات الدخان المعدلة وراثيا transgenic plants بأحتوائها على جين الغلاف البروتينى لفيروس تبرقش أوراق الدخان tobacco mosaic virus coat protein تكتسب مناعة ضد الأصابة بالفيرس (TMV). وبعد ذلك لوحظ أن هذه الظاهرة تعتبر عامة فى الكثير من الكائنات التى درست مثل الفطريات ونبات الأرابيدوبسس والدودة *C. elegans* و الدروسوفيلا والثدييات، حيث عرفت بتسميات مختلفة ففى الفطريات Fungi عرفت بأسم quelling وفى النباتات بأسم post transcription gene silencing [PTGS] وفى الديدان باسم RNAi . وخلال السنوات العشر الماضية تعرفنا على نظام مشابه يتحكم أيضا فى الإسكات الجينى ولكن بواسطة جزيئات صغيرة تعرف بأسم micro RNA [miRNA]. وتتفق كل تلك الظواهر فى القدرة على إسكات الجينات ويتم ذلك من خلال الخطوات التالية:

- بناء جزيئات من الـ RNA مزدوج الأذرع يسمى ds RNA بواسطة الأنزيم RNA dependent RNA polymerase ويعرف إختصارا (RdRP)،

وهو قد يشابه فى عمله انزيمات بناء الفيروسات النباتية (viral replicases). ويمثله الناتج الجينى لجينات مثل *SDE-1* فى الأرابيدوبسس أو *EGO-1* فى الديدان *C. elegance*.

- يقوم أحد انزيمات الهدم الـ RNA الداخلى (endonuclease) من عائلة RNase III ، *E. coli* ، أطلق عليه اسم Dicer) بهدم الـ dsRNA إلى قطع صغيرة حجمها يتراوح بين ٢١- ٢٥ نيوكلوتيدة، لها امتداد من الطرف 3`-OH نيوكلوتيدتين (3` n: overhang in each sense and antisense-). وهذا الأنزيم يحتاج فى عمله إلى طاقة ATP وقطع مزدوجة من الـ RNA لا يقل طولها عن 39 bp ، و قد يتماثل هذا الأنزيم مع الناتج الجينى *RDE-4* الخاص بـ C. elegance.
- يقوم أنزيم من نوع الـ Helicase (غير معروف خواصه على وجه الدقة) بفك ارتباط ذراعي الـ ds RNA وتحويله إلى قطع مفردة من الـ (si RNA) short interfering RNA . ويعرف هذا المعقد الإنزيمى اختصارا بـ RISC للمصطلح RNA-induced silencing complex.
- تتعرف قطع الـ siRNA مع التتابعات المناظرة لها على المرسال المستهدف والمراد إسكاته (target mRNA)، وعن طريق البلمرة polymerization أو اللصق ligation تتكون جزيئات مزدوجة يتم هدمها بسرعة بواسطة الـ Dicer.

أما فى حالة الدروسوفيلا والثدييات ومنها الأنسان، فإن ميكانيكية حدوث الأسكات الجينى بتداخل الـ RNA مازالت غير واضحة. ولكن بينت التجارب أن ارتباط الـ si RNA مع المرسال المستهدف لايحتاج لأنزيم RdRP وان النهايات 3`-OH لهذة القطع الصغيرة من الـ si RNA لا تعمل كبادئات للبناء (primers) كما فى الأنظمة الأخرى. والشكل (٣- ٩) يلخص هذه الخطوات.

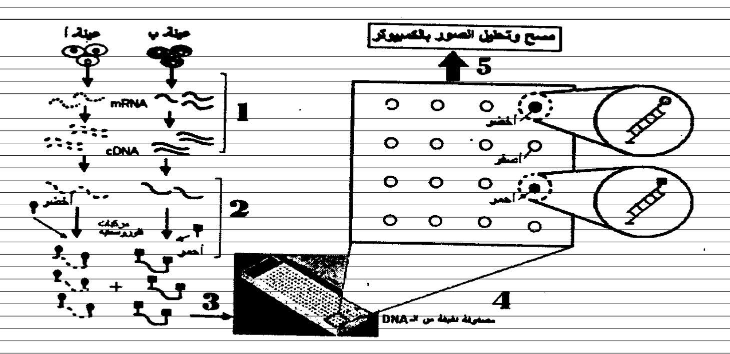


شكل (٣-٩) : رسم توضيحى لخطوات الإسكات الجينى بواسطة RNAi.

٦.٣. المصفوفات الجينومية Genomic Micro-arrays.

اصبح من الواضح لدينا ان الكائنات على إختلاف انواعها واشكالها تحتوى على أعداد كبيرة من الجينات، حيث يتم التعبير عن نسبة عالية منها فى الزمان والمكان المحدد من حياة الكائن. هذا القدر من العمل يستوجب درجة عالية من التنسيق فيما بينها، ولكن وللأسف فإن الأساليب التقنية التى كانت متاحة من قبل كانت لاتسمح إلا بدراسة التعبير الجينى لكل جين على حدة وبالتوالى، حيث كانت الصورة الشاملة للتعبير الجينى غير متاحة. ولكن فى السنوات العشر الماضية ظهرت تقنية جديدة تتيح دراسة التعبير الجينى لمجموعة من الجينات (قد تصل لعدة آلاف فى المرة الواحدة). هذه التقنية الجديدة عرفت بعدة مسميات، فأطلق عليها أسم مصفوفات الـ DNA الدقيقة DNA microarray و سميت بعد ذلك "رقائق الـ DNA " DNA chips لكونها دقيقة جدا وتشابه تكنولوجيا صناعة الدوائر المدمجة integrate circuits فى الإلكترونيات مع التأكيد على انها لا تمت لعلوم الإلكترونيات بأى صلة. وتفصيلا يمكن القول، بوجود

نوعين من تلك المصفوفات، الأول هى المصفوفات الكبيرة (نسبيا) DNA macroarrays وهى التى تتكون من نقاط من عينات الـ DNA تزيد عن ٥٠٠ ميكرون لكل نقطة – فى هذه الحالة يمكن قراءة الصور بماسحات scanners عادية، والمصفوفات الدقيقة DNA microarrays حيث لايزيد حجم العينة عن ٢٠٠ ميكرون وهنا يجب قراءة الصور الناتجة بماسحات خاصة. ومن التسميات الشائعة أيضا - الرقائق الجينية gene chips وقد استعمل هذا المصطلح فى الكثير من البحوث المنشورة فى الدوريات العلمية المتخصصة، إلا ان هذا الأسم يعتبر ملكية ادبية وتجارية لشركة Affymetrix Inc. لتسجيلها منتجا تجاريا بهذا الأسم، مما دفعها للتقدم بعدة دعاوى قضائية ضد من يستعمل هذه التسمية دون إذن مسبق منها. وبناء على هذا فقد اخترت اسم الرقائق الجينومية genomic chips لتسمية هذه التقنية دون غيرها من المسميات. شكل (٣- ١٠) يوضح أهم الخطوات لتنفيذ هذه التقنية.



شكل (٣- ١٠) : رسم توضيحى للخطوات الرئيسية لدراسة التعبير الجينى من خلال استعمال المصفوفات الجينومية.

يمكننا ان نلخص الخطوات المتبعة فى تقنية المصفوفات الجينومية فى خمس اقسام رئيسية والمرقمة فى شكل (٣- ١٠)، وهى كالتالى :

١. تحضير عينات الـ DNA تحت الدراسة، أو ما يسمي بالـ DNA المستهدف - "target DNA" — عادة يستخلص الـ RNA الكلى من عينات مختلفة تمثل ظروفا تجريبية مختلفة ومنها تعزل الرسالات mRNAs كيمائيا، وعن طريق النسخ العكسي reverse transcription يتم بناء جزيئات الـ cDNA المكملة، والتى تمثل المعين النسخي transcriptomea لهذه العينات.

٢. تعليم labeling للجزيئات المستهدفة بواسطة أصباغ فلوروسنتية fluorescent dyes وعادة تستعمل مشتقات السيانين مثل Cy3 (خضراء اللون) و Cy5 (حمراء اللون).

٣. تنقل الجزيئات المستهدفة والمعلمة بالأصباغ الفلوروسنتية لتهجينها hybridization مع أحد المصفوفات الجينومية بما تحمله من جزيئات الـ DNA العيارية أو القياسية أو ما يعرف بأسم probe DNAs، عن طريق تتابع الدورات المتحكم فيها من الـ denaturation/annealing بأستعمال المحاليل ودرجات الحرارة المناسبة. والفحص الدقيق يوضح أن الأقترانات بين الجزيئات المستهدفة (الملونة إما احمر او اخضر) سيتم مع جزيئات الـ probe المكملة لها وغير المعلمة، وعليه فى حالة أرتباط مع المصفوفة فأن اللون الأخضر سيحدد جينات العينة الأولى بينما اللون الأحمر سيحدد جينات العينة الثانية أما اللون الأصفر فسيحدد أماكن الأرتباط المشترك لجينات العينة الأولى والثانية معا، أما إنعدام اللون فسيمثل عدم وجود تهجين (نشاط جينى)، كما هو مبين بالرسم المكبر (اقصى اليمين) فى شكل (٣- ١٠).

٤. تبدو هذه الخطوة وليست فى سياق التتابع لهذه التقنية وهذا صحيح فيجب أن تكون سابقة لبداية العمل، ومع هذا يجب التعريف بها — ألا وهى صناعة المصفوفة ذاتها. تصنع المصفوفات أو الرقائق من الزجاج عادة أو من مواد بلاستيكية شفافة فى حالات نادرة بعد تغطيتها بطبقة رقيقة من مادة poly-L-Lysine وتوضع عليها بترتيب دقيق عينات الـDNA العيارى probes، بأستعمال طرق أوتوماتيكية robotic لوضع (طبع) قدر ضئيل ومحدد من هذه الـ probes على هيئة نقاط وهذه الطريقة مماثلة لطبع الدوائر الإلكترونية بما

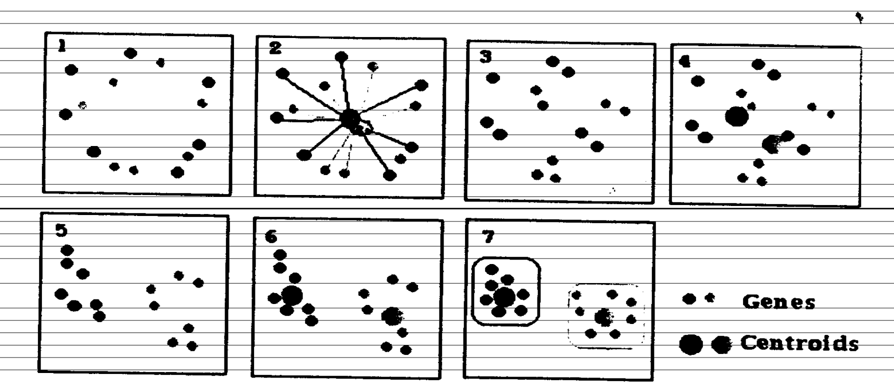
يسمى photolithography. ويتراوح عدد هذه الـ probes بين المئة حتى عشرات الآلاف، قد يمثل كل منها تتابعات صغيرة أو cDNA (جين) أو كروموسوم ونأمل أن تمثل الرقائق فى المستقبل القريب جينوما بأكمله.

٥. الخطوة الأخيرة فى هذه التقنية غاية فى الأهمية - وهى مسح وتحليل الصور الناتجة من تلك التجارب وهى تحتاج إلى بعض التفصيل.

٣.٦.١. مسح وتحليل المصفوفات Scanning & Analysis of Arrays

عادة يتبادر إلى أذهاننا فى الدراسات المعتمدة على إستعمال المصفوفات الجينومية – سؤالين – الأول يتسائل عن شدة التعبير لكل جين من جينات هذه المصفوفة تحت الدراسة – والتساؤل الثانى يتمحور حول تقدير العلاقات بين قدر التعبير الجينى لجينات المصفوفة تحت الدراسة. الأجابة على السؤال الأول عادة سهلة وممكنة، فبتوفر الأجهزة الحديثة التى تعتمد على تكنولوجيا الليزر يمكن إستحساس excitation وتقدير قدر الأمتصاص الضوئى optical absorption بدقة متناهية عند أطوال الموجات للأصباغ الفلوروسنتية المستخدمة وربطها بقدر النشاط الجينى لكل منها. أما الأجابة على السؤال الثانى فهى التى ستحتاج منا جهدا وتدقيقا أكبر من خلال إستعمال الأساليب الحسابية والإحصائية، وغالبا ما تكون غير قاطعة بل ستكون مقبول فقط عند درجة من درجات الأحتمال.

فى مبدأ الأمراتبعت طرق حساب " التجميع العنقودى" للبيانات cluster analysis لتقدير العلاقات بين مجاميع الجينات النشطة تحت ظروف التجربة. فبرامج الكمبيوتر وطرق حساب هذا التجمع لها عدة طرز ولكن أكثرها شيوعا وابسطها هى طريقة *k-means* clustering حيث يتم أولا أعطاء كل جين فى المجاميع النشطة قيمة رقمية ثم تحدد المتوسطات (مركز الثقل – centroid) لكل مجموعة ثم خطوة خطوة تجمع كل مجموعة متشابهة التعبير مع بعضها البعض وصولا للفصل مابينهم ما امكن. شكل (٣-١١) يمثل رسوما توضيحية لخطوات طريقة *k-means* clustering.



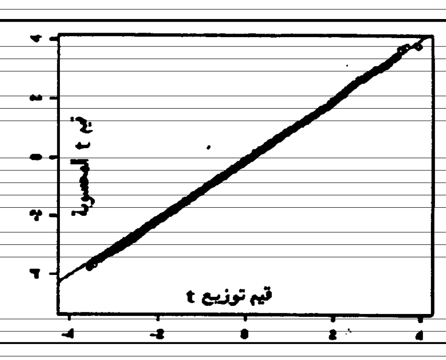
شكل (٢-١١) : رسوم توضيحية لبعض من خطوات طريقة التجميع العنقودى *k-means* clustering لتقدير درجات التشابه فى التعبير الجينى لجمعتين من الجينات النشطة المميزة بالنقاط الصغيرة الرمادية والسوداء، بينما مركز الثقل لكل مجموعة ممثل بالنقاط الكبيرة.

طرق التجميع العنقودى على إختلاف أشكالها وصورها لاشك لها أهمية تاريخية لكن الآن لا يفضلها عدد كبير من الدارسين بل يتجنبوها لعدم دقتها ويفضلون اللجوء إلى لطرق الإحصائية التقليدية للمقرنة بين مجاميع البيانات مثل اختبار t، كما هو مبين بالمعادلة التالية:

$$t = (x_{i,1} - x_{i,2}) / ((s_i + s_0)/2).$$

حيث $x_{i,1}$ هو متوسط تأثير الجين i فى المجموعة ١ و $x_{i,2}$ متوسطه فى المجموعة و s_i لخطأ القياسى داخل المجاميع و s_0 متوسط توزيع الخطأ القياسى فى كل المجاميع. عادة لايكون اختبار t فى حد ذاته كافيا لذلك وجب أختبار قيم t المحسوبة منسوبتا لتوزيع قيم t، حيث عادة يلاحظ أن قيم t لا تظهر أختلافات بين الجينات فى المجاميع وأنها تتبع توزيع منتظم ممثل بخط مائل بدرجة ٤٥، اما الجينات المختلفة فستنحرف عن هذا التوزيع بدرجات متفاوتة، كما هو مبين بالعلاقة البيانية التالية:





ومع ان اختبار توزيع t يعتبر مقبولا بشكل أولى إلا أن الطرق الأحصائية المستعملة لدراسة الإختلافات فى تجارب المصفوفات الجينوميه اصبحت اليوم أكثر تعقيدا ودقة، وبعض من تلك الطرق الإحصائية الشائعة سنتناولها بقدر من التفصيل فى الباب القادم.

٤. الأساس الحسابي للمعلوماتية الحيوية

Mathematical Bases of Bioinformatics

إعداد: أحمد المتيـنى

يمكن القول بأن المعلوماتية الحيوية هى التطبيق المباشر لتكنولوجيا المعلومات (information technology) فى مجال البيولوجيا الجزيئية (molecular biology) حيث تتم الاستعانة بأساسيات علوم الرياضة التطبيقية (applied math) والكمبيوتر (computer) والإحصاء (statistics) لتخزين وتبويب وتفهم المعلومات البيولوجية. فكم المعلومات المنتجة يوميا فى مجال البيولوجيا الجزيئية والجينومكس فاق الحدود المتوقعة وأصبح الإلمام بها أو حتى ببعضا منها بالجهود الذاتية من شبه المستحيل، فعلى سبيل المثال قدرت الأوراق العلمية المنشورة فى هذا المجال فى خلال السنوات القليلة الماضية بنحو ٢٪ من جملة البحوث المنشورة فى مجال الطب والبيولوجى عموما وذلك حسب إحصائيات قاعدة بيانات المكتبة الأمريكية القومية للطب (NLM) التابعة للمعهد القومي للصحة (NIH) والمعروفة اختصارا باسم "PubMed" [www.nlm.nih.gov]. كذلك نشرت صحيفة "الإيكونمست" فى سنة ١٩٩٩ مقالة لـ Anthony Kervalage، أحد كبار بحاث شركة Celera الشهيرة، قدر فيه حجم المعلومات التى يمكن أن ينتجها معمل مختص بهذا المجال بحوالى 100 Gb يوميا !!

إن كمية المعلومات المنتجة فى هذا المجال ضخمة للغاية ومهولة لذلك وجب الاستعانة بالكمبيوتر لأنشاء قواعد للمعلومات (databases) تبوب وتخزن بها تلك المعلومات ثم تستغل برامج الكمبيوتر (software) للبحث وتحليل وإستخلاص المعلومات، سوف نتناول بعض من هذه المواضيع بشكل أوسع فى الأبواب القادمة. أن تناولنا لهذه المعلومات من خلال مسح قواعد المعلومات المتاحة واسع ومتعدد حيث يتميز بالديناميكية العالية مقارنتا بغيره من فروع العلم، لكن يمكننا أن نلخص أهم الأهداف فى النقاط التالية:

- البحث الإستخلاصي (التجميعي) Annotation search ، حيث تستعمل مفاتيح لتصفح قواعد المعلومات باستعمال كلمات فاتحة (keywords) أو أسماء العلماء (authors) أو المواضيع المرتبطة، بغرض زيادة أو تجميع معلومات عن موضوع ما مثل معرفة – ماهو الجين (الجينات) المسئول عن مرض cystic fibrosis فى الإنسان؟
- البحث عن المثيل أو النظير Homology (similarity) search ، حيث يتم حصر وتصفح قواعد المعلومات المتاحة بغرض التعرف على نماذج متماثلة لما لدينا من معلومات مثل معرفة – هل تتابعات الجين المعرف لدينا تناظرها أو تماثلها تتابعات فى كائن آخر؟
- البحث عن المميزات Pattern search ، حيث يتم حصر قواعد المعلومات بحثاً عن أجزاء مميزة للجزيئات قد تستعمل فى معرفة وظائفها مثل – ما هي التتابعات المميزة لجين ما تعمل كمناطق ارتباط عوامل البدأ (initiation) أو الحس (enhancing) والإسكات (silencing) خلال النسخ (transcription) الجينى.
- التنبؤ Prediction ، استغلال المعلومات المتاحة بقواعد المعلومات للتوصل لإستنتاجات جديدة مثل التنبؤ بالتراكيب الثانوية (secondary structure) لبروتين ما من بيانات تركيبه الأولى (primary structure).
- المقارنات Comparisons ، استغلال قواعد المعلومات المتاحة لمقارنة تتابعات معلومة من(الأحماض النووية أو البروتينات) بغيرها من التتابعات مثل- البحث عن العائلات الجينية (gene families) فى جينوم ما.

خلال عمليات الفحص والتنقيب والإستخلاص لقواعد المعلومات هذه لابد لنا من ان نعتمد على مبادئ علم الإحصاء وبرامج و حسابات (algorethem) الكمبيوتر، وفيما يلى سنحاول الألمام ببعض من تلك المواضيع ذات الأرتباط المباشر بمجال المعلوماتية الحيوية دون توسع و الذى ليس مجاله هنا.

١.٤. نظرية الإحتمالات Probability theory.

تلعب الإحتمالات دورا هاما فى تفسير سلوك الكائنات الحية (كما أكدنا سابقا)، ويفيد الإحتمال هو ببساطة الإمكان المقرون ببعض الثقة وأيضا ببعض الشك فى وقوع حدث (event) معين. فقد تكون الثقة معدومة فى وقوع حدث ما ومع ذلك قد يفترض أنه محتمل الوقوع، ومن ناحية أخرى قد تكون الثقة فى حدوث حدث تصل لحد التأكد ومع ذلك يكتفى بذكر أن الحدث محتمل الوقوع. وعلى ذلك يشمل هذا المعنى العام للإحتمالات كل من الأحتمالات الضعيفة والقوية على حد السواء. وقد يكون الحدث بسيط (simple) أو مركبا (compound)، أى يتكون من جملة أحداث بسيطة.

الاحتمال الرياضى للحدث البسيط هو عباره عن عدد مرات نجاحه أو فشله منسوبة للعدد الكلى لعدد مرات النجاح والفشل. وهومقياس موضوعى مستقل تماما عن الهوى الشخصى. فإذا ما رمزنا للنجاح بـ a وللفشل بـ b فإن أحتمال نجاحه $(p) = a/a+b$ و أن احتمال فشله $(q) = b/a+b$، ومجموع احتمالات النجاح والفشل تساوى الواحد الصحيح $p + q = 1$. ومن هذه المعادلة يمكن إستخلاص الأساسيات التالية: $p = (1 - q)$ و $q = (1 - p)$

وفى حالة الأحداث المركبة، إذا كانت الأحداث البسيطة المكونة لها غير مستقلة عن بعضها، بمعنى إذا وقع أحدها منع وقوع الحدث أو الأحداث الأخرى فى هذه الحالة تجمع احتمالات الأحداث الفردية، أى

$$P = p_1 + p_2 + p_3 + \dots\dots\dots + p_n.$$

وإذا كان الحدث المركب مكونا من أحداث بسيطة مستقلة عن بعضها البعض، أى إذا وقع حدث منها لا يؤثر فى وقوع الأحداث الأخرى، فى هذه الحالة تضرب احتمالات الأحداث البسيطة، أى

$$P = p_1 \times p_2 \times p_3 \times \dots\dots\dots\dots \times p_n.$$

لتقدير الاحتمال الحسابى يجب أن نلم أيضا ببعض المعلومات عن التباديل (permutations) والتوافيق (combinations). التباديل - هى عدد الترتيبات المختلفة الممكن تكوينها من أحداث مختلفة عن بعضها البعض، فمثلا الحروف الثلاث a و b و c

يمكن ترتيبها فى ٦ طرق مختلفة (تباديلها): abc, acb, bac, bca, cab, cba أى أن التباديل = ١×٢×٣، فإذا كان لدينا عدد من الأحداث قدره n فإن التباديل = ١× ٢ × ٣ × .... × n أى يساوى مضروب n (n- factorial – n!). وإذا ما كان لدينا ثلاث احداث مثل a و b و c واردنا معرفة عدد التوافيق الممكنة (إذا ما اخترت أثنين أثنين) فستكون التوافيق الثلاث الممكنة ab, bc, ac ويمكن حساب التوافيق من المعادلة العامة:

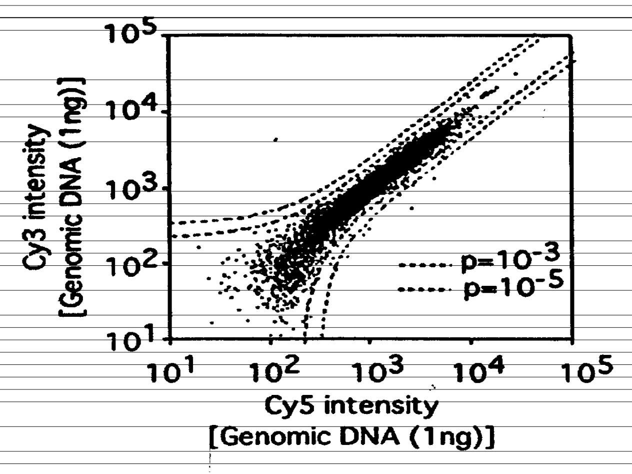
$$^{n}C_{r} = \frac{n!}{(n-r)! - r!}$$

حيث أن n = عدد الأحداث و r = عدد التوافيق المطلوبة.

٤.١.١. قواعد عامة عن توزيعات الإحتمالاتProbability distributions.

فيما يلى بعض من الحقائق والقواعد الإحصائية الحسابية التى قد تكون مفيدة فى الدراسات البيولوجية عموما وفى الجينومكس والمعلوماتية الحيوية خصوصا:

- أكثر التوزيعات إستعمالا لقياس حدود الثقة هى المنحنى الطبيعى Normal distribution والذى يعرف ايضا بأسم Gaussian's distribution وذلك للبيانات المستمرة وإستعمال مثل تلك التوزيعات يمكن الباحث من تحديد حدود الثقة لبياناته وتحديد الأخطاء التجريبية عند مستويات الإحتمال المطلوبه، كما هو مبين بشكل (١-٤) لبيانات الـ DNA microarray المتحصل عليها فى إحدى الدراسات



شكل (١٤) : توزيع بيانات الـ DNA microarray بين حدود الثقة عند إحتمال P<0.001 و P<0.00001 بإستعمال قياسات المنحنى الطبيعي.

- يستخدم توزيع الـ hypergeometric distribution فى حساب الإحتمالات عند الأختيار من أحداث مستقلة لا تعوض(without replacement) ، لهذا الغرض قد تستعمل المعادلة التالية:

$$P = 1 - \sum_{x=0}^{k-1} \frac{\begin{bmatrix} k \\ x \end{bmatrix} \begin{bmatrix} N-k \\ n-x \end{bmatrix}}{\begin{bmatrix} N \\ n \end{bmatrix}} \quad \text{where} \quad \begin{bmatrix} n \\ k \end{bmatrix} = \frac{n!}{k!\ (n-k)!}$$

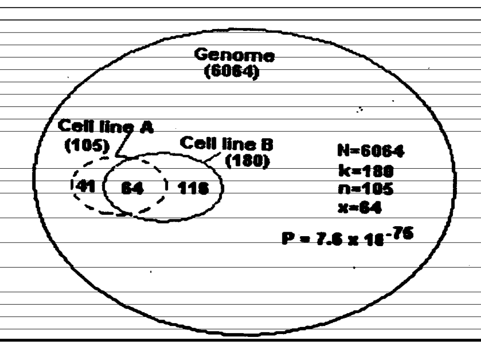
N **= total population**

k**= number of successes in total population**

n**= size of sample population**

x**= number of successes in sample population**

وتطبيق هذه المعادلة يمكن تمثيله بتقدير المعنوية لبيانات الـ microarray لأثنين من السلالات المختلفة لنفس الكائن، كما هو موضح بشكل (٤-٢).



شكل (٤-٢) : رسم توضيحى Venn diagram لمقارنة بيانات الـ DNA microarray لاثنين من السلالات لنفس الكائن (نفس الجينوم).

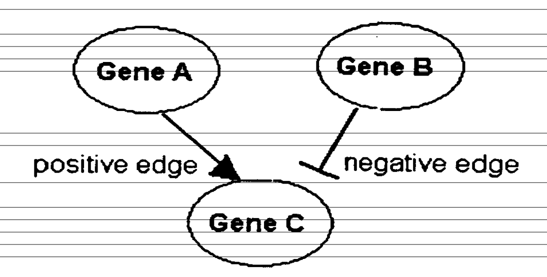
٤. ٢. طريقة بياس الإحصائية Bayesian statistics.

اقترح عالم اللاهوت والرياضة الأنجليزى Thomas Bayes طريقة إحصائية لتقدير الإحتمالات تختلف عن الطرق الأخرى، فى كونها تعتمد فى تقدير الإحتمال أعتمادا على تقديرات سابقة مستقاه من النتائج التجريبية ذاتها، وفى حينه لم تلق تلك الحسابات التأييد المطلوب من الرياضيين التقليديين، ولكنها فى السنوات الأخيرة زاد الأهتمام بها لمحاولة تطبيقها على بيانات الوراثة الجزيئية والجينومكس، حيث تعرف الآن بأسم طريقة بياس Bayesian statistics. وتعتمد الطريقة على تقدير الأرجحية العظمى maximum likelihood estimate (MLE) وذلك للمقارنة بين نظريتين فرضيتين hypothesis لتقدير ارجحية أحدهما، بأستعمال العلاقة الآتية:

$$\ell = \frac{L(x \setminus H_0)}{L(x \setminus H_1)} \;\&\; \chi^2 = -2\ln\ell = -2\ln\left[\frac{L(x \setminus H_0)}{L(x \setminus H_1)}\right]$$

حيث H_0 تمثل الفرضية الأولى و H_1 تمثل الفرضية الثانية،وتقاس المعنوية بإختبار مربع كاي χ^2.

يعتبر الوراثيون ان هذه الطريقة مناسبة لكونها بسيطة فى تمثيل الجينات كذلك يمكن ان تعكس العلاقات لتأثير الجينات على بعضها البعض، والأستفادة من المعلومات السابقة فى وضع الفرضيات وكذلك لقدرتها على التعامل بكفاءة مع البيانات المتداخلة مثل بيانات DNA microarrays. وقد انبثق عن هذه الطريقة طريقة تسمى Bayesian network حيث تمثل العلاقات بصورة وصفية qualitative أولا ثم تدرس كميا quantitative ثانيا، وهذا الأسلوب يستخدم فى الوراثة بكثرة. فعلى سبيل المثال يمثل الجين بإنتفاخ node ومن حوافه تخرج أسهم التأثير فإذا كان موجبا مثل برأس السهم وإذا كان التأثير سلبيا مثل بالخط المنتهى كما هو موضح بشكل (٤-٢).



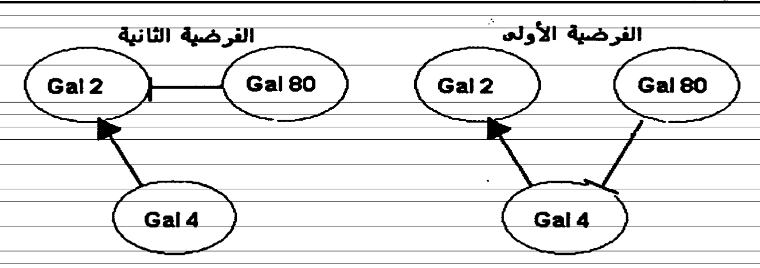
شكل (٤-٢) : تمثيل للعلاقة بين الجينات بناء على Bayesian network.

والرسم السابق يمثل الجزء الوصفى للطريقة حيث مثلت النظرية الفرضية بناء على النتائج التجريبية، أما الشق الكمي للطريقة فيعتمد على تقدير الإحتمالات لمدى تأثير النشاط الجينى لكل من الجينين A و B على نشاط الجين C سلبا أو إيجابا، كما هو موضح بالجدول التالى:

الجين A	الجين B	احتمال التأثير الإيجابى على الجين C	احتمال التأثير السلبى على الجين C
+	+	0.6	0.4
-	+	0.01	0.99
+	-	0.99	0.01
-	-	0.4	0.6

والبيانــات الموجــودة بالخانــة المظللــة مــن الجـدول الســابق تعنــى أن زيــادة نشــاط up-regulation لكل من الجينين A و B سيؤدى إلى زيادة نشاط الجين C (up-regulated) بأحتمال قدره ٦٠٪ وباحتمال قدره ٤٠٪ فإن زيادة نشاط A و B ستؤدى إلى خفض نشاط الجين C (down-regulation).

كذلك تستخدم هذه الطريقة فى تفهم طبيعة تأثير الجينات على بعضها البعض، ففـى دراسـة قـام بهـا .Hartemink et al سـنة ٢٠٠١ باسـتغلال بيانـات ٥٢ محاولـة للـ DNA microarrays لمجموعة من الجينات المتداخلة فى التـأثير علـى تمثيـل الجلاكتـوز Galactose regulatory network، وفى جزء محدد من تلك الدراسـة (علـى سـبيل المثـال) اقترحوا فرضيتين لتأثير ثلاث من هذه الجينات، كما هو مبين بالشكل (٤-٤).



شكل (٤-٤) : فرضيات تأثيرات الجين Gal 80.

ولإختبـار ارجحيـة أى مـن النظـرتين الفرضـيتين السـابقتين تسـتخدم Bayesian network statistics عـن طريـق حسـاب معامـل أو قيـاس يسـمى Bayesian score (M)، يحسب من العلاقة الآتية:

$$M = \log [P(M/D] = \log [P(M)] + \log [P(D/M)] + c.$$

حيـث M - النمـوذج (الفرضـية) و D - نتـائج الـ microarray المتحصـل عليهـا و c - ثابت. وبحسابات الأرجحية maximum likelihood method تحسب المعاملات لكلا الفرضـيتين، فوجـد أن هـذا المعامـل للفرضـية الأولى يسـاوى 44.0 – بينمـا كـان المعامـل للفرضية الثانيـة يسـاوى 34.5 –، وعليـه يمكـن الاسـتنتاج بـأن الفرضـية الثانيـة أكثـر ارجحية لتفسير النتائج المتحصل عليها.

ولكن أتضح أن Bayesian statistics ليست مناسبة للعديد من الحالات لذلك تستعمل طرق إحصائية أخرى مثل برامج ماركوف الإحصائية لهذا الفرض.

٤.٣. برامج (نماذج) ماركوف Markov Models .

هذا النوع من البرامج والطرق الإحصائية هو الأكثر إستعمالا فى مجال المعلوماتية الحيوية لكونها أكثر ملائمة لأهداف تلك الدراسات، ولكن الأساس الرياضى الإحصائى لهذه البرامج غاية فى التعقيد وتحتاج المتخصصين فى الرياضة التطبيقية وعلوم الكمبيوتر للتعامل معها. هذه الحقيقة تستدعى لفت الأنظار لأهمية الفريق البحثى فى هذا المجال الحديث من المعرفة، ففى مجال الجينومكس والمعلوماتية الحيوية يجب أن يتعاون الوراثيون مع المتخصصين فى مجالات عديدة مثل الكيمياء الحيوية Biochemistry والفزياء الحيوية Biophysics وعلماء الرياضة Mathematicians وغيرهم لإنجاز الأهداف المرجوة. بالنسبة لنا نحن البيولوجيون لايهمنا التبحر فى الأسس الرياضية لهذه البرامج، بل يهمنا تفهم الإمكانيات والأهداف العامة لتلك البرامج لتطبيقها فيما يخصنا من دراسات، وحمدا لله أن هناك العديد من برامج الكمبيوتر software التى يمكن أن تساعدنا فى هذا التطبيقات.

وفى مجال الجينومكس والمعلوماتية الحيوية تستغل برامج ماركوف الإحصائية لتحديد أرجحية التوافق matching أو عدم التوافق mismatching أو إمكانية الحذف deletion أو الأضافة insertion لتتابعات من الأحماض النووية أو البروتينات لتحديد مصادرها أومعرفة أسلافها ancestors المحتملة. وهذه البرامج الإحصائية تقلد العقل الإنسانى فى تعامله مع حالات عديدة مثل إلتباس الألفاظ عند أحد المتكلمين، فقد يكون ذلك لإختلاف اللهجات أو للإصابة بألتهاب فى الحنجرة أو لعيب خلقي فى المتحدث يجعل من مخارج الألفاظ غير واضحة أو لعدم التمكن من اللغة وغيرها، ولكن عادة، المستمع يتوصل لفهم هذا الحديث حيث يقيس فى عقله هذه الألفاظ المبهمة على نماذج لكلمات قريبة منها فى النطق ومنها يختار الأقرب للمعنى، فمثلا فى حديث ما، قال المتحدث فى سياق كلامه " المؤمن كيس فطن" وطبعا لايستقيم المعنى فالمقصود (كما يعيه أغلبنا دون عناء) هو " المؤمن كيّس فطن".

ودون الدخول فى التفاصيل الحسابية يمكن تشبيه برامج ماركوف هذه بلعبة game من العاب التسلية (تعتمد على الحظ المحسوب) تستعمل فيها رقعة من المربعات مثل رقعة الشطرنج أو لعبة السلم والثعبان مثلا، ولكنها لعبة غريبة ومملة ! وقواعد اللعبة يمكن أجمالها فى :

(١) كل مربع يوفر لك مجموعة من الرموز أو الحروف (فى حالة الـ DNA فلديك ٤ حروف وفى البروتينات لديك ٢٠ حرفا) وتكرارات أو نسب هذه الحروف لبعضها البعض متغير (مثلا فى الـ DNA الحرف A له نسبته والحرف G له نسبته كذلك الحال مع الحرفين C و T)، المهم أن كل مربع يعطى نسب مختلفة عن المربعات الأخرى (فمربع به نسب عالية من A ونسب منخفضة من الثلاث الباقية والمربع المجاور يعطى نسب عالية من G ونسب منخفضة من الثلاث الباقية وهكذا.....).

(٢) داخل كل مربع، كل حرف له قيمة حسابية تتراوح بين الواحد الصحيح والصفر تميزه بناء على تكراره.

(٣) فى نهاية كل دور من أدوار اللعبة يحسب مجموع الدرجات الكلية إلتى حصلت عليها من مجمل إنتقلاتك بين المربعات فيما قد يسمي بالنتيجة النهائية final score، والذى يكسب الدور هو من يجمع أكثر النقاط (ليس دائما).

(٤) إذا انتقلت من مربع إلى آخر لابد أن تنال عقوبة penalty (خصما من الدرجات)، لكن هناك مربعان بالرقعة — مربع الحذف deletion ومربع الأضافة insertion – يمكنك الأنتقال إليهما فى أى وقت دون عقوبات لكن لن تحصل على أى درجات منهما. وإذا ما كانت هذه قواعد اللعبة، فإن الهدف من هذه اللعبة هو التداول لتتابع ما من الـ DNA لمعرفة ارجحية إنتمائه لتتابع أخر عن طريق جمع أكبر قدر من النقاط مع تجنب (بقدر المستطاع) العقوبات.

يبدو لى أن مفهوم اللعبة قد أتضح بعض الشيء، ولكن ليس بالقدر الكافى لذلك دعنا نلعب دورا مبسط فى هذه اللعبة الغريبة. فإذا ما كان لدينا تتابع من ٢٥ نيوكلوتيدة من الـ DNA على النحو التالى:

```
    5    10 11   15   20   25
AAATATTATACTATCGGCAGCGCG
```

وإذا كان المتاح فى هذا الدور من اللعبة مربعين أثنين فقط، كما هو مبين على النحوالتالى:

مربع (١)	مربع (٢)
توفر بقدر كبير كل من T's و A's ولكن يسمح بالحرفين G's و C's بندرة	توفر بقدر كبير كل من G's و C's ولكن يسمح بالحرفين T's و A's بندرة

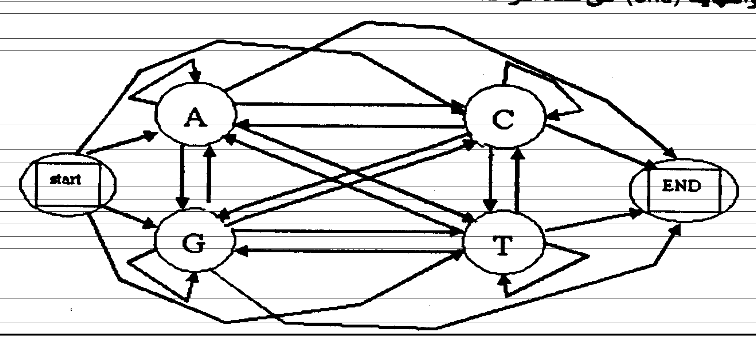
لكسب هذا الدور (بمعنى معرفة الترتيب المعطى)، يمكنك أن تبقى فى المربع الأول حيث تتلقى دائما درجات عالية عن الحروف من الأول حتى الحرف رقم ١٥ فكلها من الحروف A او T كمواصفات المربع، الأستثناء هو الحرف رقم ١١ فهو C حيث سيكون لك الخيار فى أن تنتقل للمربع الثانى لتحصل على درجة عالية عن الحرف C ولكنك ستنال عقوبة بالخصم للأنتقال ولكن ستجد أن الحرف التال فى الترتيب رقم ١٢ هو T فتنتقل مرة ثانية للمربع الأول للحصول على الدرجة العالية لكنك ستنال عقوبة الأنتقال مرة ثانية، وعند بلوغ الحرف رقم ١٥ تنتقل للمربع الثانى حيث معظم الأحرف من رقم ١٥ إلى رقم ٢٥ من النوع C او G، تقابلك نفس الحالة عند الحرف رقم ٢٠ حيث أنه A (يمكن تسمية هذه الطريقة بالإستراتيجية الأولى). أما الإستراتيجية الثانية هى أن تبقى فى المربع الأول من الحرف رقم ١ حتى رقم ١٥ دون تحرك متحملا الدرجة المنخفضة للحرف رقم ١١ لكن متجنبا عقوبة الخصم مرتين، ومن الحرف رقم ١٥ حتى النهاية فتنتقل للمربع الثانى متحملا الدرجة المنخفضة عند الحرف رقم ٢٠ ولكن متجنبا الخصومات. ويتضح ان الإستراتيجية الثانية هى التى ستعطيك أعلى الدرجات لكسب الدور.

فيما سبق حاولنا تبسيط الأمر بقدر المستطاع لتفهم المهام المطلوبة من برامج ماركوف الإحصائية، لكن مع الأخذ فى الاعتبار أن رقعة المربعات ستكون من عدد كبير من المربعات وعليها أن تختار المسار path الأمثل بينها للوصول للنتيجة المرجوة، وتخيل حجم التباديل والمحاولات الواجب القيام بها لإتمام الدور. ومما زاد من الطين بله أن

برامج ماركوف هذه يجب أن يكون بها قدر من المعلومات الخفية على اللاعب، لذلك تسمى برامج ماركوف الخفية Hidden Markov Models وتعرف إختصارا HMM. ومثال ذلك انك ستبدأ اللعب دون أن تعرف مكان الأبتداء أو مكان الإنتهاء، كمن يلعب أحد ألعاب التسلية (السلم والثعبان مثلا) فى الظلام وأن شخصا آخر يلعب نيابة عنه دون أن يطلعه على ما يجرى من أحداث وعليه جمع النقاط فقط !!!

٤.٢.١. برامج ماركوف و الوراثة Markov Models & Genetics .

لدراسة تتابع ما من الـ DNA باستعمال برامج ماركوف، يجب أن نحدد رموز (حروف) الموضوع تحت الدراسة، وهو فى هذه الحالة معروفة لدينا، فهى النيوكلوتيدات الأربع (A,G,C,T) التى يمكن أن تتواجد أو تتبادل مواقعها على هذا التتابع، وشكل (٤-٥) يوضح تلك الرموز والتبادلات الممكنة لكل منها (قدرها ١٦ لكل منها)، ومع تجاهل البداية (start) والنهاية (end) فى هذه المرحلة.



شكل (٤-٥) : رسم توضيحي للنيوكلوتيدات الأربع واحتمالات التبادل والإحلال بينها.

ومن معلوماتنا الوراثية نعلم أن الجينومات فى الكائنات الراقية تتميز بوجود جزر من تتابعات المتكررة من C $_p$ G (C $_p$ G islands) ولكن نعلم أيضا أن الجينوم يتعرض لعمليات الـ methylation الدائمة التى تحول السيتوسين إلى الثيمين أى أن مثل تلك الجزر ستصبح T $_p$ G islands، كما نعلم أن مناطق البروموتورات (promoters) لا تحدث فيها عمليات الـ methylation ، أى تبقى بها الجزر من C $_p$ G دون تغير. وحيث

أن احتمال كل رمز فى موقع ما S_p سيتوقف فقط على احتمال الرمز الذى قبله فى الموقع p-1 أى S_{p-1}، وباستعمال تتابعات معروف أنها غنية فى جزر C_pG (overabundant model) و تتابعات أخرى معروف أنها فقيرة فى تلك الجزر (underabundant model)، يمكننا حساب الاحتمالات لتباديل الرموز بينها البعض وذلك بناء على المشاهدات التجريبية، والجدول التالى يلخص تلك الاحتمالات من نتائج النموذج الفقير فى تلك الجزر:

From / To	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

والجدول التالى يلخص الاحتمالات المتحصل عليها من نتائج النموذج الغنى بتلك الجزر:

From / To	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

مما سبق من بيانات، أصبح لدى البرنامج نموذجين للقياس، الأول نموذج يمثل التتابعات الفقيرة فى تلك الجزر (underabundant model)، والثانى نموذج يمثل التتابعات الغنية فى هذه الجزر (overabundant model). وحيث أن إحتمال حدوث رمز S عند الموقع P (S_p) يتوقف فقط على الرمز الذى قبله عند الموقع P-1 (S_{p-1}) وليس على جملة التتابعات السابقة، ولحساب إحتمال أن تتابع ما يتوافق (fits) مع نموذج ما يقدر حاصل ضرب الاحتمالات الشرطية (conditional probabilities) كالآتى :

$$.P(x) = P(x_L/x_{L-1})\ P(x_{L-1}/x_{L-2})\ ....\ P(x_2/x_1)\ P(x_1)$$

والتى تمثل رياضيا بالمعادلة التالية:

$$P(x) = P(x_1)\prod_{i=2}^{L} a_{x_{i-1}x_i}$$

حيث $a_{x_{i-1}x_i}$ هو قيمة الإحتمال لإنتقال الرمز من الموقع $i-1$ إلى الموقع i.

ولسهولة الحسابات دعنا نقول بأن الاحتمالات المحسوبة من النموذج الفقير فى C p G [underabundant model] =

P(A) = P(T) = 0.3 & P(C) = P(G) = 0.2

وبأن الاحتمالات المحسوبة من النموذج الغني فى C p G [overabundant model] =

P(A) = P(C) = P(G) = P(T) = 0.25

الآن دعنا نحاول دراسة التتابع GGCGACG على سبيل المثال لتحديد مع أى من النموذجين يتوافق. واحتمال هذا التتابع تبعا للـ [underabundant model] هو:

P(G)P(G/G)P(C/G)P(G/C)P(A/G)P(C/A)P(G/C)

والذى يساوى-

(0.20)(0.298)(0.246)(0.078)(0.248)(0.205)(0.078)

= 0.000000453499

واحتمال هذا التتابع تبعا لـ [overabundant model] يساوى -

(0.25)(0.375)(0.339)(0.274)(0.161)(0.274)(0.161)(0.274)(0.274)(0.125)

= 0.0010526

وبناء على تلك الحسابات - يمكن القول بأن هذا التتابع أكثر ارجحية فى ان يكون منتمى للنموذج الثانى أى غني فى C p G ويتبع الـ overabundant model.

ويتضح لنا من حسابات الإحتمالات السابقة، أن القيم تؤول للصفر (zero) بسرعة كبيرة وللتغلب على ذلك يستعمل التحويل اللوغارتمى أو مايسمي بـ log statistics.

فى إستخدامنا السابق لطريقة ماركوف كنا نبحث عن التوافق أو الإنتماء، ولكن يمكن أن تستخدم هذه الطريقة أيضا للتفرقة (discrimination) بين المعطيات. فمثلا إذا ما تساءلنا عن مدى الإختلاف بين النموذجين السابقين أى الـ overabundant و الـ underabundant. وإذا ما كانا غير مختلفين بالقدر الكافى عن بعضهما البعض، فلن تتوفر لدينا معلومات حقيقية وكافية لتقدير أرجحية أن تتابع ما فى أن ينتمي لأى منهما. ولإختبار قدر الاختلاف بينهما تستخرج قيم اللوغارتم للأساس الثانى لنسب الاحتمالات للنموذجين من الجدولين السابقين وتحسب القيم كما هو فى الجدول التالى :

From / To	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

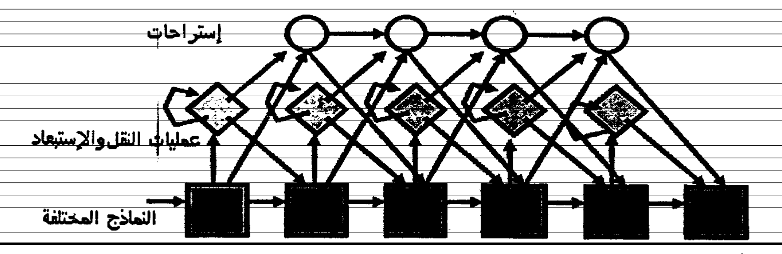
ومن القيم السابقة يتضح أن الفرق واضح بين كلا النموذجين فالقيم التى ستنتمى للنموذج underabundant model ستكون سالبة القيمة أما القيم التى ستنتمى إلى النموذج overabundant model ستكون موجبة، مما يؤكد أنهما نموذجين مختلفين.

٢.٣.٤. نماذج ماركوف الخفية Hidden Markov Models

نماذج ماركوف التى تناولناها بالشرح والتفسير فيما سبق محدودة الإستعمال، وعادة تكون متطلبات البحث والتدقيق فى المعلومات الوراثية هوتداول أكثر من نموذج أو ما يسمى بالوضعية (state)، فلمعرفة أرجحية إنتماء بروتين ما لأحد العائلات البروتينية (protein families) نحتاج أكثر من نموذج قد يصل عددها إلى عشرة، كذلك تكون التتابعات تحت الدراسة أكبر بكثير من تلك التى إستعملناها فى أمثلتنا السابقة. ففى حالة النموذجين (الغنى و الفقيربالجزر من C p G)، لكى نبحث عنهما فى مقطع كبير من الـ DNA الجينومى، فلن نعرف أماكن تلك الجزر على وجه التحديد (أى خفية علينا)،

لذلك وجب دمج النموذجين مع بعضهما مع السماح للبرنامج بالتنقل خلال مسار محدد (path) وعادة يكون خفي لحساب احتمالات التحول (transition) واحتمالات الإستبعاد (emission) لكل حدث أو رمز.

والتخطيط المبين بشكل (٤-٦) يبين تصميم لبرنامج إحصائى من نوع ماركوف الخفي (HMM)، حيث يتم تحديد النماذج القياسية المختلفة التى يمكن قياس التتابع تحت الدراسة عليها، وبرامج تنفيذ عمليات المسح من تحولات (transitions) و إستبعادات (emission)، أما مناطق الأستراحة بين التحولات والإستبعادات لمسار (path) ما من ضمن المسارات المحتملة فيجب تواجدها كمناطق محايدة لسهولة تحديد المسار الأمثل.



شكل (٤-٦) : رسم تخطيطى لتصميم أحد برامج ماركوف الخفية (HMM).

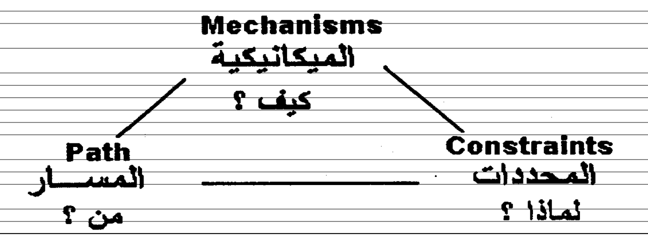
ويتوفر على شبكة المعلومات العالمية INTERNET وعلى الموقع الخاص بمدرسة الطب بجامعة واشنجتون (Washington University in St. Louis, School of medicine) عدة برامج لإستعمال برامج ماركوف الخفية لدراسة تتابعات البروتينات والمعروفة بأسم HMMER وآخر نسخة منها صدرت سنة ٢٠٠٣ باسم hmmer 2.3.2

<u>الموقع:</u> http://hmmer.wustl.edu

ومع عظمة هذه الأساليب الحسابية الحديثة وإسهاماتها الجليلة فى تفهم مواضيع مثل البيولوجيا الجزيئية والجينومكس إلا أن المبدأ الفيزيائي العام، ألا وهو "عدم التأكد - Uncertainty principle " يمكن أن يلقي بظلاله عليها، <u>بالاعتقاد بأن</u>

محاولة التنبؤ بالسلوك البيلوجي أوالوراثي في الطبيعة هي عملية غير مؤكدة بل مستحيلة !!

دعنا الأن نتناول موضوعا وراثيا مرتبط بموضوع دراستنا وقد يقع تحت طائلة مبدأ عدم التأكد هذا، ألا وهو دراسة التطور بين الكائنات إعتمادا على دراسة مدى التشابه والإختلاف بين البروتينات لكائنات مختلفة تربطها درجات من القرابة، أو ما يسمى «بفيلوجينيا» البروتين (protein phylogeny). ان دراسة تتابعات الأحماض الأمينية للسلاسل الببتيدية في الكائنات المختلفة لإستخلاص العلاقات الفيلوجينية منها يعتمد على ثلاثة محاور رئيسية متداخلة : أولها هو ميكانيكيات (mechanisms) التغيير لدراسة أسباب منشأ التغيرات بين البروتينات (تحاول الأجابة على كيف؟ حدث التغيير) و ثانيها هو المسار (path) الذي سلكته، لدراسة مسار التغير بين المتشابه والمتباعد (تحاول الأجابة على ممن؟ جاءت الإختلافات) وثالثها هو المحددات (constraints) لهذه التغيرات لدراسة العوامل التي قد تساعد أوتعاكس ظهور هذه الإختلافات (تحاول الأجابة على لماذا؟ توجد الإختلافات). والرسم التوضيحي التالي يبين هذه المحاور الثلاث.



فكل محور منها يؤثر على الآخر بدرجة أو بأخرى، ولأكتمال الدراسة ولزيادة مصداقيتها لابد من تناولها جميعا في الحسبان عند دراسة فيلوجينية بروتين ما ضمن عدد من الكائنات، ولكن القصور المادى لتصميم التجارب يحول دائما دون هذا. فعند دراسة الفيلوجيني أو المسار فلا بد أن نفترض نموذجا ثابتا للمحورين الآخرين، مثل إفتراض ميكانيكية محددة (من ضمن احتمالات عديدة) وكذلك يجب إفتراض وجود نموذجا للإنتخاب محدد (من ضمن إحتمالات عديدة). بناء على هذا التبسيط التجريبي (القصور) فإن النتائج التي نتوصل إليها عادة غير مؤكدة وبالتالي تندرج تحت طائلة مبدأ

عـدم التأكـد لــ Heisenberg. لتوضيح هـذه النقطـة دعنـا نقـارن بـين تتابعـات الأحماض الأمينية لثلاثيات الببتيدات الأفتراضية التالية من أربعة كائنات مختلفة:

. . VGM . .

. . VAM . .

. . VPM . .

. . VLM . .

فالإختلافات عند الموقع الأوسط يمكن تفسيرها تبعا لواحد من الأحتمالات الثلاث التالية:

- قـد يكـون الكـودون الخـاص بـالحمض الأوسـط يمثـل نقطـة سـاخنة (hot spot) للطفور. هنا ركزنا على "الميكانيكية" دون المسار والمحددات.
- قد يمثل ذلك خطوات المنشأ من سلف مشترك بالترتيب التالى G > A > P > L حيـث أن هـذا الأحتمـال ينـتج مـن إسـتبدال نيوكلوتيـدة واحـدة بينمـا بـاقى الأحتمـالات تنشـأ مـن إسـتبدال أكثـر مـن نيوكلوتيـدة واحـدة. هنـا ركزنـا علـى "المسار" دون الميكانيكية والمحددات.
- قـد يمثـل ذلـك عـدم وجـود محـددات (مثـل الإنتخـاب الطبيعـى) عنـد الكـودون الأوسط. هنا ركزنا على "المحددات" دون الميكانيكية والمسار.

٤.٤. ملحق (1) Appendix.

البرمجـة (computer programming) تعتـبر الآن جـزء أساسـى مـن المعلوماتيـة الحيويـة، وإن كانـت ليسـت مـن مهـام البيولـوجيين وتعتمـد علـى المتخصصـين فـى علـوم الكمبيوتر، إلا أن هذا الملحق يتناول بشكل سريع بعض من المعلومات الأساسية لأحد لغات البرمجة للكمبيوتر، وهى لغة Perl ، الأكثر إستعمالا فى البرمجة الوراثية وعسى أن تكون النواة ليزداد أهتمام أحد البيولوجين بالموضوع ثم التخصص فى هذا المجال الجديد.

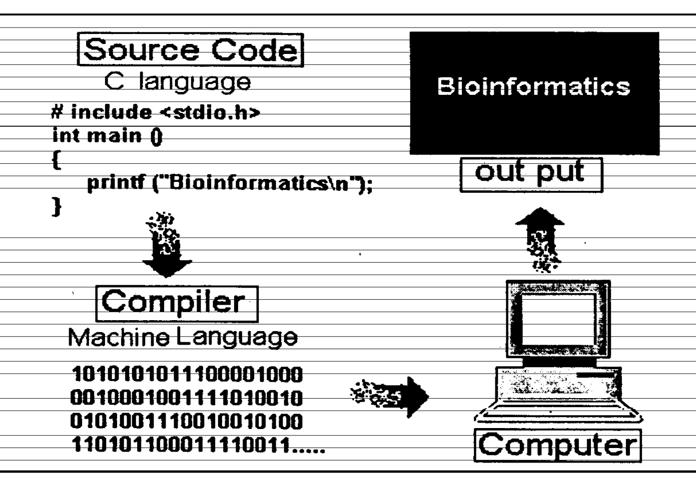
لغات البرمجة للكمبيوتر عديدة وفى كل يوم هناك الجديد فى هذا المجال، وعلى تعدد واختلاف تلك اللغات فلكى يتفهمها الكمبيوتر فيجب أن تحول جميعها إلى لغة واحدة خاصة بالكمبيوتر وتعرف بأسم بلغة الآلة (machine language) وهى لغة ثنائية (binary) كما هو موضح بشكل (٤-٧). ولغات البرمجة أو ما يعرف بأسم source codes متعددة الأنواع ويمكن إجمالها فى الآتى:

١. لغات المعالجة (compiled languages) – مثل لغات C, C++, Java وهى تمتاز بسرعة العمل والتنفيذ ولكنها صعبة فى الكتابة وتستخدم فى برامج أنظمة التشغيل (operating systems) وبرامج معالجة الكلمات (word processing).

٢. لغات النصوص (scripting language) – مثل لغات Perl, Pythom, Ruby, Tel - وهى تمتاز بسهولة الكتابة ولكن بطيئة فى العمل والتنفيذ لذلك فهى مناسبة للبرامج الصغيرة.

٣. لغات الشبكة العالمية للمعلومات (Web languages) – مثل html, PHP, Java script – وهى خاصة بالربط بين صفحات الشبكة المتداخلة وربطها مع قواعد المعلومات.

٤. لغة قواعد المعلومات (databases language) – مثل SQL - لغة قادرة على تداول المعلومات بين قواعد المعلومات المختلفة وشبكة المعلومات العالمية



شكل (٤-٧) : تخطيط يمثل خطوات تداول برنامج مكتوب بلغة C وتحويله للغة ثنائية binary حتى يتفهمه الكمبيوتر وينفذه.

واللغة الأكثر إستعمالا فى مجال المعلوماتية الحيوية هى لغة Perl وهى إختصار لـ Practical Extraction and Report Language وقد اقترحها Larry Wall لأنها مناسبة لتداول النصوص وإستخلاصها من شبكة المعلومات العالمية والقدرة على تلقى وإرسال المعلومات بسهولة وأن كان يعيبها بطء التنفيذ. وفيما يلى بعض من القواعد الأساسية لهذه اللغة:

قواعد بناء الجمل Syntax rules:

1- Statements are terminated by a semi-colon
   - Print ("Hello!\n");
2- Text blocks are determined by curly brackets
   - If ($a ==$b) {
```
        print ("a=b!\n");
     }
```

3- Comments are indicated by sharp sign

$a = 10; # set $a equal to 10

4- Separate variable names with spaces, otherwise space has no meaning

- $a + $b; is the same as $a + $b;

5- Common conventions: \n = new line, \t = tab, " " = string

6- **Order matters:** statements are evaluated in descending order.

المهام Assignments :

1- Equal sign represent variable assignment

- $A = B

2- Binary assignments operators:

- $A = $A + 5; => $A += 5;
- $B = $B – 6; => $B -= 6;

انواع المتغيرات Variable types :

1- Dollar-sign ($) variable represents a scalar (number) or string

- $DNA_length = 10;
- $DNA_sequence = "ATTAGCCGAT";

2- At-sign (@) variable represents an array, ($) sign represents individual array element

- @DNA = (A,T,T,A,G,C,C,G,A,T):
- $DNA[0] is equal to "A"; $DNA[1] is equal to "T";

3- Percent-sign (%) represents a hash, ($) represent individual hash element

- %DNA = ("First" => "A", "Second" => "T");
- $DNA["First"] is equal to "A";

العمليات الحسابية والمنطقية Arithmetic & Logical operations :

1- Addition: $a = 5 + 6; # $a equal 11
2- Subtraction: $a = 6 – 5; # $a equal 1
3- Multiplication: $a = 3 * 2; # $a equa; 6
4- Division: $a = 6 / 2; # $a equal 3
5- Modules: $a = 3 % 2 = 5
6- Auto-increment: $a = 0; $a++; # $a is equal 1
7- Identify test: if ($a == 6); # $a is equal to 6
8- Not equal test: if ($a != 6); # $a is not equal to 6
9- Less than, greater than: $a < 6; $b > 5
10- AND operator: $a == 6 && $b $b > 4; # $a equal to 6 AND $b is greater than 5
11- OR operator: $a == 6 D $b < 4; # $a equal to 6

الأوامر الشرطية Conditional statements :

1- if/else statements:
   - if statement {do if statement is true}
   - else statement {do if statement is false}

الأوتار Strings :

1- String connection is a period (.)
   - $a = "Hello"; $b = "World";
     $c = $a. $b; # $c is "Hello World"
2- String length: $length = length($string)
   - $text_length = length($c); # $text_length is 10
3- String reverse: $rev_string = reverse($string);
   - $rev_c = reverse($string); # $rev_c is "dlro WolleH"

الحلقات Loops :

1- for statements:

```
- for($i = 0; $i < 20; ++) {
  $DNA[$i] = "A";
}
```

2- foreach statements:

```
- foreach $i (@DNA) {
      $i = "A";
}
```

3- while statement:

```
- while ($i < 20) {
  $DNA[$i] = "A";
      $i++;
}
```

المعطيات والمخرجات Input/Output :

1- Standard input: <STDIN>
2- Standard output: print
3- Opening file: open (FILEHANDEL,"filename");
   - open (DNASEQ, "dnaseq.txt file");
4- Reading file: single line => $DNA = <FILEHANDEL>;
   whole file => @DNA = <FILEHANDEL>;

المصفوفات Array :

1- Split function: splits a string into an array of letters.

```
$seq = "ATAGCCAT");
@DNA = split(//,$seq);  # $DNA[0] is "A", $DNA[1] is "T"
```

2- Push/Pop: Push adds value to end of array, pop removes last value of array.

```
push (@DNA, "G");  # @DNA is {A,T,A,G,C,C,A,T,G}
$last = pop (@DNA); # $last is "G"
```

3- Reverse: reverses order of the array

```
@DNA = reverse (@DNA); # @DNA is {T.A,C,C,G,A,T,A}
```

4- Length of array: scalar @array

```
$size of array = scalar @DNA; # $size of array is 8.
```

تعبيرات عامة Regular Expressions :

1- Matching: $string =R/pattern/

```
$DNA = "ATATAAAGA";
if ($DNA  =R/TATA/ {
        print ("Contains TATA element\n");
}
```

2- Substitution: $string =Rs/pattern/replacement pattern/(g);

3- $DNA =Rs/TATA/GGGG/g; # $DNA is now "AGGGGAAGA"

4- Wildcards: [ATGC] matches A or T or G or C;

[^0-9] matches all non-digit characters;

A{1,5) matches a stretch of 1 to 5 "A" characters

فيما سبق حاولنا التعرف على أهم القواعد للغة Perl ولمزيد من التفصيل يمكن اللجوء إلى المواقع المتخصصة في هذه اللغة كالتالية:

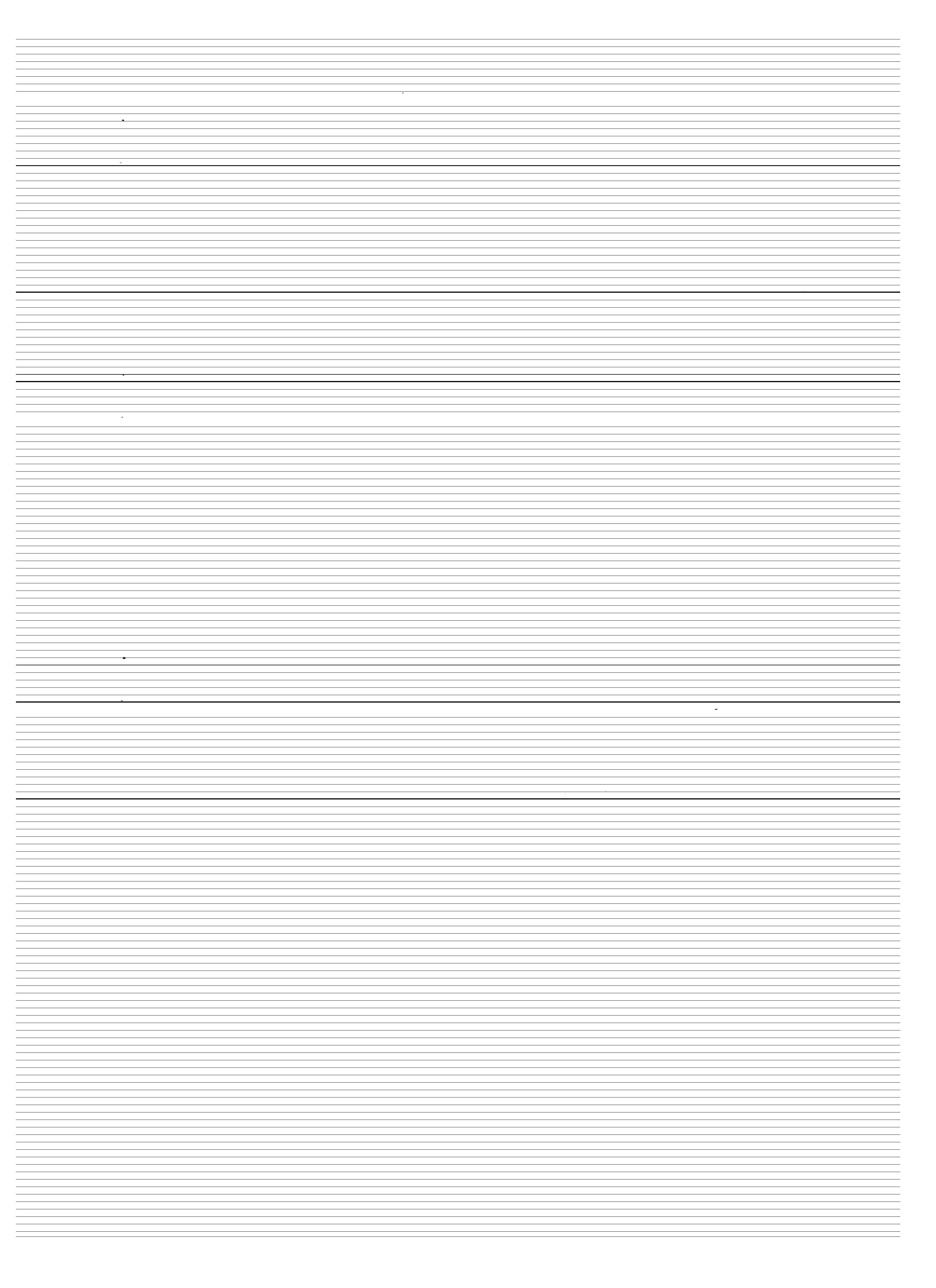
www.perl.com/perl

www.perl.com/cpan

مثال: البرنامج الأساسي لدراسة توافق تتابع ما بلغة Perl.

```
for (i = 0 to l = m – n) {
      j = 0;
      while (j < n and Qj == Ti + j) {
            if (j == n – 1) {
                  return l;
            }
            j = j + 1;
      }
}
```

where: Q = DNA sequence (Query)
T = Chromosome sequence (Target)
Q_j = Nucleotide in Query sequence at position j
T_i = Nucleotide in Target sequence at position i
n = length of Q
m = length of T



٥. قواعد المعلومات البيـــولـــوجيـــة

Biological Databases

إعداد : آمال عبد العزيز

تعتمد المعلوماتية الحيوية اعتمادا جوهريا على توفر قواعد المعلومات البيولوجية بمفهومها الحديث. وقاعدة المعلومات فى أبسط صورها قد تكون ملفا صغير يحفظ عدة بيانات عن موضوعا ما أو قد تكون عبارة عن مخزن كبير (أرشيف) لحفظ ملفات عن قطاع من المواطنين و الذى قد يشغل عدة أدوار من مبنى كبير، ومع أن هذه المحفوظات هى صورة من صور قواعد المعلومات، إلا أنها ليست المقصودة فى مجال اهتمامنا الحالى. فقواعد المعلومات (databases) التى نقصده هنا هى : مجموعة البيانات أو المعلومات المخزنة والمبوبة بطرق إلكترونية باستعمال تقنيات الكمبيوتر وتتميز بمدخل امامى (front-end) حيث يمكن للمستعمل الدخول والبحث والتدقيق بالقاعدة، ويجب أن يتوفر لها أيضا مدخل خلفى (back-end) حيث يمكن للمشغل أو المسئول عن القاعدة أن يتناول بيانات القاعدة بالإضافة و الحذف أو التحديث (up-dating). ويرجع الفضل إلى مهندس الكمبيوتر الشهير Charles Backman فى تصميم أول تلك القواعد بمفهومها الحديث مع مطلع الستينات من القرن الماضى. ويمكن أعتبار الأطلس الذى نشرته العالة Dayhoff Margaret وزملائها سنة ١٩٦٥ بعنوان "Atlas of protein sequences and structure"، والذى كان النواة التى أصبحت فيما بعد قاعدة المعلومات عن البروتينات والمعروفة بأسم PIR database وهو قد يعتبر أول تلك المحاولات الحديثة.

وقواعد المعلومات البيولوجية بالنسبة للمهتمين بالمعلوماتية الحيوية توفر لهم المعلومات الأساسية والتى بدونها تصبح الدراسة مستحيلة، ويمكن إجمال فوائد قواعد المعلومات البيولوجية فى الآتى :

(١) توفير المعلومات الجينية والبروتينية للباحثين.

(٢) تخزين وتبويب المعلومات الجينية والبروتينية على هيئة صفحات مقروءة بالكمبيوتر.

٥. ١. طرز قواعد المعلومات البيولوجية Types of biological databases

عادة تقسم قواعد المعلومات البيولوجية لطرز عديدة تتبع فيها معايير مختلفة، وفيما يلي أهم الطرز حسبا لتلك المعايير:

(١) بناء على نوع الكائن - فقد تقسم القاعدة بناء على تخصصها عن معلومات لكائن واحد فقط، فقد تختص قاعدة بجينات الإنسان وأخرى بجينات الفأر وثالثة بجينات نبات الأرز وهكذا.

(٢) بناء على نوع البيانات – وهناك قواعد تتميز بنوع البيانات التى تختص بها، فهناك قواعد لتتابعات النيوكلوتيدات و أخرى لتتابعات البروتينات وثالثة لبيانات الأبعاد الثلاثية (3D) للجزيئات ورابعة للتعبير الجينى وهكذا.

(٣) بناء على طبيعة البيانات – هل البيانات فى القاعدة تمثل النتائج الأولية (القواعد الأولية) أم تمثل النتائج بعد التحليل والتدقيق (القواعد الثانوية).

(٤) بناء على الإتاحة (Availability) - تقسم القواعد أيضا على مدى إتاحتها على الشبكة الدولية للمعلومات، فمنها ما هو متاح للجمهور مجانا ومنها ما هو متاح مجانا مع الحفاظ على حقوق الملكية الفكرية ومنها ما هو متاح للمتخصصين أو المشتركين فقط وهكذا.

(٥) بناء على التصميم – يعتبر التصميم الفنى للقاعدة هو أهم المعايير التى تميز قاعدة عن أخرى، وفيما يلى نبذة مبسطة عن نظم تصميم وبرمجة قواعد المعلومات.

٥.١.١. نماذج تصميم قواعد المعلومات Modeling Databases.

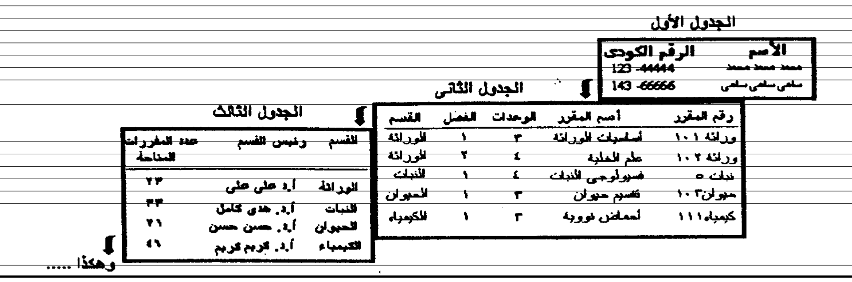
يعتبر تصميم وبرمجة قواعد المعلومات من المهام الأساسية لإنشاء تلك القواعد، وهناك العديد من الطرق وبرامج الكمبيوتر المتاحة لهذا الفرض، ولكن أهم ميزتين يجب أن يتصف بهما أى نموذج لتصميم قاعدة ما، هما:

- البساطة Simplicity.
- تجنب تكرار البيانات Redundancy.

وفى مبدأ الأمر كانت النماذج بدائية لا تختلف كثيرا عن الجداول أو ما يسمى spreadsheet ويطلق على هذه القواعد النماذج المسطحة (ذات البعدين) flat-files. والجدول التالى يوضح أحد تلك النماذج لتسجيل بيانات الطلبة.

الأسم	الرقم الكودى	المقرر	التقدير
محمد محمد محمد	٤٤٤٤٤	وراثة (١٠١)	ممتاز
محمد محمد محمد	٤٤٤٤٤	وراثة (١٠٢)	جيدجدا
محمد محمد محمد	٤٤٤٤٤	وراثة (١٠٣)	جيدجدا
سامى سامى سامى	٦٦٦٦٦	وراثة (١٠١)	جيدجدا
سامى سامى سامى	٦٦٦٦٦	وراثة (١٠٢)	جيد
سامى سامى سامى	٦٦٦٦٦	وراثة (١٠٣)	ممتاز

ويلاحظ أن مثل هذه القواعد به كمية كبيرة من البيانات المتكررة والبحث فيها يستغرق وقتا كبيرا، لذلك لجأ المصممين لتجزئة البيانات وتقليل التكرار عن طريق تصاميم سميت بالنماذج الهرمية أو الشجرية Hierarchical (tree) model حيث ترتب جداول المعلومات فى أنظمة هرمية مرتبطة ببعضها البعض كما هو موضح بالجداول التالية:



الجدول الأول

الأسم	الرقم الكودى
محمد محمد محمد	123 -44444
سامى سامى سامى	143 -66666

الجدول الثانى

رقم المقرر	أسم المقرر	الوحدات	الفصل	القسم
وراثة ١٠١	أساسيات الوراثة	٣	١	الوراثة
وراثة ١٠٢	علم الخلية	٤	٢	الوراثة
نبات ٥	فسيولوجى النبات	٤	١	النبات
حيوان١٠٣	تقسيم حيوان	٣	١	الحيوان
كيمياء١١١	أحماض نووية	٣	١	الكيمياء

الجدول الثالث

القسم	رئيس القسم	عدد المقررات المتاحة
الوراثة	أ.د علي علي	٢٣
النبات	أ.د. هدى كامل	٣٣
الحيوان	أ.د. حسن حسن	٢١
الكيمياء	أ.د. كريم كريم	٤٦

وهكذا .....

حيث يقل تكرار البيانات بالجدول وتوفر معلومات أكبر ولكن مثل هذه التصاميم قل استعمالها الأن، وتصمم القواعد الأن بأستخدام برامج الكمبيوتر إعتمادا على الأسس الرياضية والجبرية مثل نماذج Relational model التى تعتمد على لغة SQL وهى اختصار Structured Query Language وهى لغة من لغات الكمبيوتر توفر طرق البحث والتدقيق والتحديث لقواعد المعلومات. وحديثا تستعمل لغات أخرى لإنشاء قواعد المعلومات خصوصا التى تتداول النصوص والوسائل المرئية والسمعية عبر شبكة المعلومات العالمية internet مثل لغات FTP, HTML, CORBA, XML وغيرها.

٥. ٢. قواعد المعلومات الوراثية Genetic Databases.

يتواجد الآن عدد كبير من قواعد المعلومات المتخصصة لتبويب وتجميع البيانات الخاصة بالأحماض النووية والبروتينات، وهناك عدد محدود منها يمثل أهم تلك القواعد وأكثرها شهرة واستعمالا. ففيما يخص تتابعات الأحماض النووية فإن القاعدة التابعة لـ NCBI الأمريكية و قاعدة EMBL الأوربية وقاعدة DDBJ اليابانية هم الأهم والأكثر استعمالا، وفيما يخص البروتينات فأن قاعدة SWISS-PROT الأوربية هى الأهم. بالإضافة لتلك القواعد فهناك المئات من القواعد الأخرى الأقل أهمية ولكنها قد تكون أكثر تخصصا، كل فى مجاله، والجدول التالى يلخص بعض من أهم تلك القواعد:

القاعدة	الموقع	ملاحظات
VectorDB	http://vectordb.atcg.com/	تحتوى على العديد من تتابعات أدوات النقل
Codon usage	http://www.dna.affrc.go.jp/	أنواع الكودونات فى الكائنات المختلفة.
TRANSFAC	http://transfac.gbf.de/	تتابعات عدد من *cis*-acting DNA تعمل كمنظمات للنسخ.
RNA world	http://www.imb-jena.de/	تحتوى على وصلات لقواعد أخرى بها عديدة من تتابعات الـ RNAs الصغيرة.
NDB	http://ndb-mirror-2.rutgers.edu/	تجمع البيانات عن الأحماض النووية.
PRINTS	http://www.biochem.ucl.ac.uk/	تشمل مجموعة من التراكيب البروتينية المحفوظة، لتعريف البروتينات.
Pfam	http://www.sanger.ac.uk/	تشمل على أجزاء بروتينية .domains
GSDB	http://www.ncgr.org/gsdb/	بها قدر ممتاز من التتابعات الجينومية.
REBASE	http://rebase.net.com/	تشمل معلومات عن الإنزيمات المحددة ومناطق تعارفها.
GCRDb	http://www.gcrdb.uthscsa.edu/	تحتوى معلومات وافية عن الـ G-proteins.

٥. ٢. ١. المركز القومى الأمريكى للمعلومات البيوتكنولوجية NCBI .

تنبهت الحكومة والكونجرس فى الولايات المتحدة الأمريكية في نوفمبر سنة ١٩٨٨ إلى الحاجة الملحة لإستخدام الكمبيوتر لتخزين وتبويب المعلومات المتنامية في مجالات الكيمياء الحيوية البحتة و الكيماء الحيوية الطبية، لذلك عمل على سن القوانين التي تساعد في عمل مكتبة قومية خاصة بالكيمياء الحيوية الطبية وأن تكون تابعة للمكتبة الطبية القومية NLM التابعة بدورها للمعهد القومى للصحة NIH، وذلك للعمل على تحسين وتوفير الأدوات التحليلية الجديدة وللمساعدة في إستيعاب المفاهيم الجديدة فى الوراثة الجزيئية و بالتالى معرفة دورها في دراسة الحالات المرضية. وعليه أنشئ المركز القومى للمعلومات البيوتكنولوجية NCBI حيث حدد له أربعة مهام رئيسية وهي :

١. ابتكار أدوات أوتوماتيكية تستطيع تحليل و تخزين المعلومات الخاصة بالبيولوجيا الجزيئية عموما و الوراثة الجزيئية و الكيمياء الحيوية خصوصا.

٢. تسهيل إستخدام البرامج التحليلية المتاحة للإتصال بالبيانات الرئيسية او الأساسية (على سبيل المثال إتاحة تبادل المعلومات بين العلماء المهتمين بالنواحي الطبية).

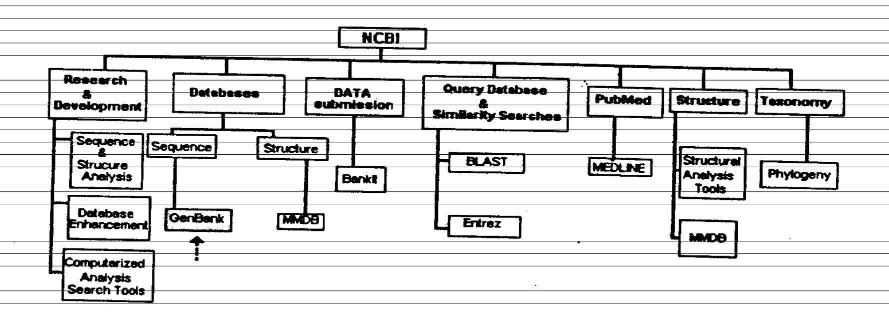
٣. العمل على ربط المجهودات العالمية مع بعضها البعض في مجال البيانات البيولوجية.

٤. العمل على تسهيل الأتصالات البحثية لتوفير وسائل التحليلات التركيبية و الوظيفية للبيولوجيا الجزيئية.

وقد بدأ الـ NCBI متواضعا ولكن مع مرور الأيام أصبح أكبر قاعدة للمعلومات في العلوم البيولوجية الجزيئية، فهو يقدم الأن خدمات كثيرة من خلال مواقع متخصصة على الشبكة العالمية للمعلومات (الأنترنت)، ويمكن إجمالها في الخدمات التالية:

1. PubMed (Public MEDLINE).
2. BLAST (Basic Local Alignment Search Tool).
3. Entrez.
4. BankIt (World Wide Web Submission).
5. OMIN (Online Mendelian Inheritance in Man).
6. Taxonomy.
7. Structure.

وشكل (١-٥) يلخص أهم تلك المهام والخدمات.



شكل (٥-١) : رسم توضيحى يمثل المهام والخدمات التى يوفرها المركز القومى للمعلومات البيوتكنولوجية NCBI.

فمن خلال الموقع Entrez أو الموقع BLAST أو حتى مباشرة من موقع NCBI يمكن الدخول إلى الموقع المعروف باسم " بنك الجينات" GenBank هو بدون شك الأهم لنا نحن الوراثيون.

٥.٢.١.١. موقع بنك الجينات GenBank.

يتضمن هذا الموقع تتابعات النيوكلوتيدات لكل من الـ DNA والـ RNA، حيث تحفظ البيانات المختصة والتى يتم تجميعها من الباحثين مباشرة أو من المعامل العلمية المختصة أو من مكاتب البراءات العلمية Patent Offices وكذلك يتم ضم المعلومات من القواعد الأوربية EMBL واليابانية DDBJ على أسس يومية كجزء من التعاون الدولى فى هذا المجال. وقد بلغت المعلومات المخزنة فى هذا الموقع فى أغسطس ٢٠٠٤ حوالى ٤١,٨ بليون قاعدة من حجم تتابعات قدره ٣٧,٣ مليون تتابع تمثل جينومات العديد من الأنواع حقيقية وغير حقيقية النواة. وتقدر الزيادة التى تمت على البنك خلال الفترة المنصرفة من سنة ٢٠٠٥ بحوالى ٧,٩ مليون قاعدة جديدة كما ذكر .Benson et al سنة ٢٠٠٥.

يمكننا دخول هذا البنك الجينى عن طريق الموقع Entrez حيث يوجد الموقع GenBank والذى بدوره تبوب فيه البيانات تحت العديد من تحت الأقسام : ومن أهمها تحت قسم التقسيم sequence-based taxonomy أى حسب نوع وأسم الكائن العلمى، حيث يحتوى الآن على حوالى ١٦٥,٠٠٠ نوع مختلف وتقدر الزيادة بهذا القسم بحوالى ٢٠٠٠ نوع كل شهر - أو من خلال تحت قسم مصدر التتابع حيث تتراكم يوميا البيانات الجديدة من مختلف التتابعات لكن تبقى التتابعات المستخلصة من EST هى الأكثر حتى الآن فهى تمثل ٢٩٪ من حجم البيانات، ولكن يمكن منها الوصول إلى بيانات الجينات المحددة UniGene والتى تبلغ حوالى ٧٠٠,٠٠٠ جين مستقل يمثلون ٥٠ كائن مختلف.

عند طلب أى معلومات من هذا الموقع GenBank عن تسلسل تتابعات جين أو قطعة جينومية، تظهر لنا صفحة البيانات كملف مستقل تترتب فيه المعلومات بنظام خاص. فمثلا عند البحث عن تتابعات جين المالتيز maltase gene فى احد أنواع الخميرة Candida albicans فسيظهر لنا الملف المبين بشكل (٥-٢)، حيث تحدد بعض من

المعلومات الأساسية، ففى السطر الأول تحت عنوان locus يبين به الرقم الكودى للمدخل XM 714334 يليه حجم التتابع 1713 bp ثم نوع الجزئ mRNA فى هذه الحالة وفى النهاية تاريخ الإدخال. وتحت definition تحدد بيانات اسم الجين والأسم العلمى للكائن المدروس، والتفاصيل التقسيمية لهذا الكائن يمكن ملاحظتها تحت عنوان organism. أما أهم المعلومات لقاعدة البيانات فتوجد تحت عنواني accession و version حيث يحدد الرقم الكودى للعينة وهو ثابت لا يتغير حتى لو تم تغير وتدقيق نفس التتابعات، لكن يظهر لنا رقم جديد للمدخل يسمى المعرف identifier وهو فى حالتنا gi 68473242 وهو لزيادة التحقق لكل قاعدة. وعادة توجد عناوين أخرى تحدد المصدر والمرجع الذى قدم المعلومات وأماكن نشرها ودرجة التحقق منها وغيرها من المعلومات، وتحت عنواني features و comment تتواجد معلومات اضافية عن الكائن والجين والحالات المشابهة، وطبيعة الجزيئات المدخلة للقاعدة وعادة تستعمل الإختصارات التالية لتميزها:

con = constructed sequence,	est = expressed sequence tag,
gss = genome survey sequences,	htg = high throughput genomic,
new = new since last release,	patent = patent, and
sts = sequence tagged site.	

وفى النهاية ترتب تتابعات الأحماض الأمينية المترجمة من هذا التتابع وبعدها تتابعات النيوكلوتيدات لهذا الجين والتى قدرها فى هذه الحالة ١٧١٣ زوج من القواعد.

LOCUS	XM_714334 1713 bp mRNA linear PLN 29-SEP-2005
DEFINITION	Candida albicans SC5314 maltase (CaO19_7668), mRNA.
ACCESSION	XM_714334 XM_435011
VERSION	XM_714334.1 GI:68473242
KEYWORDS	
SOURCE	Candida albicans SC5314
ORGANISM	Candida albicans SC5314 Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; mitosporic Saccharomycetales; Candida.
REFERENCE	1 (bases 1 to 1713)
AUTHORS	Jones,T., et al.
TITLE	The diploid genome sequence of Candida albicans
JOURNAL	Proc. Natl. Acad. Sci. U.S.A. 101 (19), 7329-7334 (2004)
PUBMED	15123810
REFERENCE	2 (bases 1 to 1713)
AUTHORS	Jones,T. et al.
TITLE	Direct Submission
JOURNAL	Submitted (16-APR-2004) Stanford Genome Technology Center, 855 California Avenue, Palo Alto, CA 94304, USA
COMMENT	PROVISIONAL REFSEQ: This record has not yet been subject to final NCBI review. This record is derived from an annotated genomic sequence (NW_139452).
COMPLETENESS:	incomplete on both ends.
FEATURES	Location/Qualifiers
source	1..1713 /organism="Candida albicans SC5314" /mol_type="mRNA" /strain="SC5314" /db_xref="taxon:237561" /chromosome="R"
gene	1..1713 /locus_tag="CaO19_7668"

```
                /db_xref="GeneID:3638946"
CDS             1..1713
                /locus_tag="CaO19_7668"
                /note="gene whose transcription is induced by
                maltose and sucrose and repressed by glucose; one
                of two genes similar to P.angusta MAL1 and to
                seven maltase and maltase-like genes in S. cerevisiae"
                /codon_start=1
                /transl_table=12
                /product="maltase"
                /protein_id="XP_719427.1"
                /db_xref="GI:68473243"
                /db_xref="GeneID:3638946"
/translation=
        "MSEHKWWKEAVVYQIWPASYKDSNGDGV……………GNY
         KLVLTNVDKDS KDALSPYEARMYVVD"

ORIGIN
atgagtgaac ……………… ccatacgagg ctagaatgta tgtagttgattaa //
```

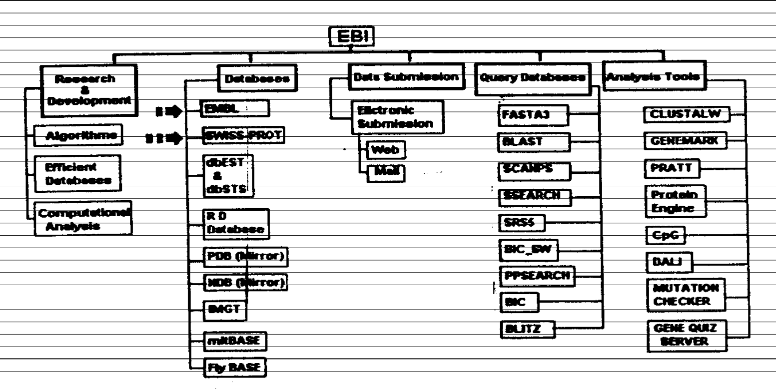
شكل (٥-٢) : صفحة من ملف فى الـ GenBank توضح تتابعات جين المالتيز فى أحد أنواع الخميرة.

٥. ٢. ٢. معهد المعلوماتية الأوربي EBI.

بدأ التفكير فى إنشاء هذا المعهد سنة ١٩٨٦ وهو نتيجة التعاون بين العديد من الدول الأوربية بغرض التخزين والتبويب الألكترونى للمعلومات الوراثية خصوصا تتابعات النيوكلوتيدات والبروتينات، وهو الإمتداد الطبيعى لعامل البيولوجيا الجزيئية الأوربيةEuropean Molecular Biology Laboratory والمعروفة إختصارا EMBL والمنتشرة فى العديد من الدول الأوربية والمقر الرئيسى للمعهد يوجد فى معمل البيولوجيا الجزيئية فى بلدة Hinxton بأنجلترا. وأهداف المعهد مماثلة لأهداف المركز الأمريكى المناظر NCBI، ويمكن تلخيصها فى الآتى:

- تقديم وتطوير تكنولوجيا تتبع المعلومات.
- تطوير وتحديث برامج دراسة تكنولوجيا المعلومات.
- توفير برامج تعليمية وتدريبية فى مجالات تكنولوجيا المعلومات.

ولتحقيق تلك الأهداف قام المعهد بإنشاء العديد من الخدمات والمواقع على شبكة الأنترنت يمكن تلخيصها فى شكل (٥-٢).



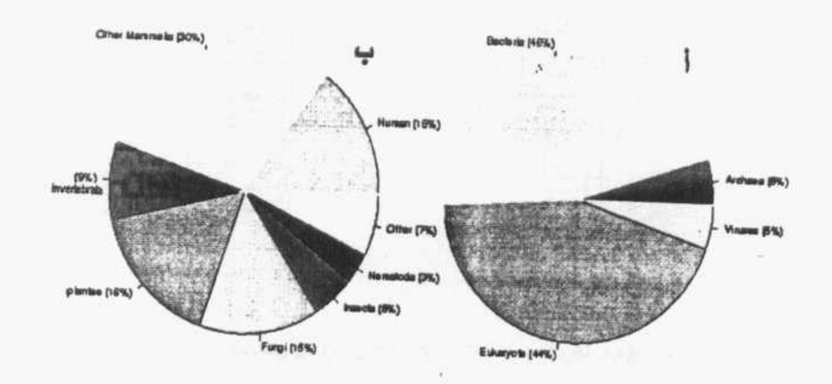
شكل (٥-٢) : رسم توضيحى يمثل المهام والخدمات التى يوفرها معهد المعلوماتية الأوربي EBI.

ومن جملة تلك المواقع والخدمات يبقى موقع EMBL الخاص بتتابعات النيوكلوتيدات والموقع SWISS-PROT لتتابعات البروتينات هما الأهم بالنسبة للوراثيين، والقاعدة EMBL لاتختلف كثيرا عن نظيرتها فى المركز الأمريكى GenBank، لذا سنكتفى هنا بالتعريف بالقاعدة SWISS-PROT

٥.٢.٢.١. قاعدة معلومات البروتينات SWISS-PROT.

تشمل قاعدة البيانات المعروفة بأسم SWISS-PROT على البيانات الخاصة بتتابعات البروتينات وقد بدأت القاعدة من قسم الكيمياء الحيوية بجامعة جينيف Geneva University (المعروف الآن بأسم المعهد السويسرى للمعلوماتية Swiss Institute of Bioinformatics) وقد إنفرد الـ EBI بأصدار هذا الموقع مع مطلع التسعينات من القرن الماضى، وعنوانه على الأنترنت

هو http://www.expasy.org/sprot/userman.html. وفى عام ١٩٩٦ اضيفت خدمه جديدة لهذا الموقع عرفت بأسم TR-EMBL وهو تحت موقع يختص بتخزين بيانات ترجمة سلاسل النيوكلوتيدات المخزنة فى EMBL اوتوماتيكيا بإستعمال برامج الكمبيوتر. وحتى سبتمبر من ٢٠٠٥ بلغ عدد المحفوظات accessions إلى 2151724 تتابع بروتينى موزعة على كافة الممالك الحية كما هو موضح بشكل (٥-٤).



شكل (٥-٤) : رسم بيانى يمثل نسب تتابعات البروتينات المخزنة بموقعى SWISS-PROT / TR-EMBL : (أ) الممالك المختلفة (ب) الكائنات حقيقية النواة.

وقواعد التبويب وترتيب المعلومات فى هذه القاعدة متماثل لحد كبير للقواعد والنظم المتبعة فى القاعدة الأمريكية السابق بيانها، وشكل (٥-٥) يمثل صفحة لأحد المحفوظات بقاعدة SWISS-PROT، وهى بيانات حول البروتين الإنسانى tyrosine protein-kinase والذى يتضح من تتابعاته أنه يتكون من ٥٣٦ حمضا أمينيا.

٥. ٢. ٣. البنك الياباني لبيانات الـ DNA (DDBJ).

البنـك اليابـاني للمعلومـات الخاصـة بتتابعـات الـ DNA هـو ثالـث القواعـد الكبيرة والأكثر إستعمالا، وهو الآخر ترجع بداياته إلى سنة ١٩٨٦ حيث أنشئ كوحدة تابعة للمعهد القـومي للوراثـة NIG بمشيما بـالقرب مـن طوكيـو تحت إشـراف وزارة التعليم والعلوم اليابانية. وما لبث أن انضم إلى التعاون الدولي في هـذا المجـال مـع EBI الأوربي و NCBI الأمريكي. وفي سنة ٢٠٠١ تم الاعتراف بهذه القاعـدة علـى أنهـا نـواة مركز للمعلومات البيولوجية وبنـك معلومـات الـ DNA الـذى يعـرف إختصـارا بأسـم (CIB-DDBJ). ويعتمد المركز على المدخلات التى يقدمها العلماء اليابانيون فـى هـذا المجال وإن كان يقبل المعلومات مـن خـارج اليابـان أيضـا. والمعهد يـدعم بشـكل مؤكـد البحوث في مجال البيولوجية الجزيئية والتتابعات الجينومية، وإن كان التعامـل معـه أكثر صعوبة لإعتماده على اللغة اليابانية.

```
ID   FYN_HUMAN      STANDARD;      PRT;   537 AA.
AC   P06241;
DT   01-JAN-1988 (REL. 06, CREATED)
DT   01-FEB-1994 (REL. 28, LAST SEQUENCE UPDATE)
DT   01-OCT-1996 (REL. 34, LAST ANNOTATION UPDATE)
DE   PROTO-ONCOGENE TYROSINE-PROTEIN KINASE FYN (EC 2.7.1.112) (P59-FYN)
DE   (SYN) (SLK).
GN   FYN.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 86287278. [NCBI, Geneva]
RA   SEMBA K., NISHIZAWA M., MIYAJIMA N., YOSHIDA M.C., SUKEGAWA J.,
RA   YAMANASHI Y., SASAKI M., YAMAMOTO T., TOYOSHIMA K.;
RL   PROC. NATL. ACAD. SCI. U.S.A. 83:5459-5463(1986).
CC   -!- FUNCTION: IMPLICATED IN THE CONTROL OF CELL GROWTH.
CC   -!- CATALYTIC ACTIVITY: ATP + A PROTEIN TYROSINE = ADP +
CC       PROTEIN TYROSINE PHOSPHATE.
CC   -!- SUBUNIT: ASSOCIATES THROUGH ITS SH3 DOMAIN, TO THE P85 SUBUNIT OF
CC       PHOSPHATIDYLINOSITOL 3-KINASE.
CC   -!- SIMILARITY: CONTAINS A COPY EACH OF THE SH2 AND SH3 DOMAINS.
CC   -!- SIMILARITY: TO OTHER PROTEIN-TYROSINE KINASES IN THE CATALYTIC
CC       DOMAIN. BELONGS TO THE SRC SUBFAMILY.
DR   EMBL; M14333; G181172; -. [EMBL / GenBank / DDBJ] [CodingSequence]
DR   EMBL; M14676; G338228; -. [EMBL / GenBank / DDBJ] [CodingSequence]
DR   PIR; A24314; TVHUSY.
DR   PIR; A26389; TVHUSR.
DR   PDB; 1SHF; 31-OCT-93.
DR        [ENTRY / RASMOL / 3D IMAGE / HSSP ENTRY / SCOP]
DR   SWISS-3DIMAGE; FYN_HUMAN.
DR   GeneCard; FYN.
DR   MIM; 137025; -.
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP.
DR   PROSITE; PS00109; PROTEIN_KINASE_TYR.
DR   PROSITE; PS50001; SH2.
DR   PRODOM [Domain structure / List of seq. sharing at least 1 domain]
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   PROTO-ONCOGENE; TRANSFERASE; TYROSINE-PROTEIN KINASE; PHOSPHORYLATION;
KW   ATP-BINDING; MYRISTYLATION; SH3 DOMAIN; SH2 DOMAIN; PALMITATE;
KW   LIPOPROTEIN; 3D-STRUCTURE.
FT   INIT_MET      0      0
FT   LIPID         1      1       MYRISTATE.
FT   LIPID         2      2       PALMITATE (BY SIMILARITY).
FT   LIPID         5      5       PALMITATE (BY SIMILARITY).
FT   DOMAIN       81    142       SH3.
FT   DOMAIN      148    245       SH2.
FT   DOMAIN      270    523       PROTEIN KINASE.
FT   MOD_RES      11     11       PHOSPHORYLATION (BY PKC) (BY SIMILARITY).
FT   NP_BIND     276    284       ATP (BY SIMILARITY).
FT   BINDING     298    298       ATP (BY SIMILARITY).
FT   ACT_SITE    389    389       BY SIMILARITY.
FT   MOD_RES     419    419       PHOSPHORYLATION (AUTO-) (BY SIMILARITY).
FT   MOD_RES     530    530       PHOSPHORYLATION.
FT   CONFLICT    183    183       A -> S (IN REF. 2).
FT   CONFLICT    434    436       A -> R (IN REF. 2).
FT   STRAND       86     88
FT   STRAND       92     92
FT   STRAND       98     98
FT   STRAND      102    102
FT   TURN        104    106
FT   STRAND      107    113
FT   HELIX       134    136
FT   STRAND      137    139
SQ   SEQUENCE   537 AA;  60630 MW;  01B18DB0 CRC32;
     GCVQCKDKEA TKLTEERDGS LNQSSGYRYG TDPTPQHYPS FGVTSIPNYN NFHAAGGQGL
     TVFGGVNSSS HTGTLRTRGG TGVTLFVALY DYEARTEDDL SFHKGEKFQI LNSSEGDWWE
     ARSLTTGETG YIPSNYVAPV DSIQAEEWYF GKLGRKDAER QLLSFGNPRG TFLIRESETT
     KGAYSLSIRD WDDMKGDHVK HYKIRKLDNG GYYITTRAQF ETLQQLVQHY SERAAGLCCR
     LVVPCHKGMP RLTDLSVKTK DVWEIPRESL QLIKRLGNGQ FGEVWMGTWN GNTKVAIKTL
     KPGTMSPESF LEEAQIMKKL KHDKLVQLYA VVSEEPIYIV TEYMNKGSLL DFLKDGEGRA
     LKLPNLVDMA AQVAAGMAYI ERMNYIHRDL RSANILVGNG LICKIADFGL ARLIEDNEYT
     ARQGAKFPIK WTAPEAALYG RFTIKSDVWS FGILLTELVT KGRVPYPGMN NREVLEQVER
     GYRMPCPQDC PISLHELMIH CWKKDPEERP TFEYLQSFLE DYFTATEPQY QPGENL
\\
```

[A sample Swiss-Prot entry.]

شكل (٥-٥): عينة من احد ملفات القاعدة SWISS-PROT.

٦- رصّ التتابعات

Sequence Alignment.

إعداد: أمـــير يـــــسن

١.٦ مقدمة.

أيا ما كان تعريف الحياة فما هى إلا تعبير عن لغة من الأحماض النووية والبروتينات وغيرها من الجزيئات البيولوجية. وهى لغة تبلغ من الثبات حدا تبدو معه التباينات الهائلة بين شتى الكائنات الحية، الدقيقة منها والراقية والمندثرة منها والباقية، وكأنها لا تعدو أكثر من مجرد لهجات محلية للغة أم واحدة. صحيح أننا نعلم اليوم أن ما يصدق لبكتيريا القولون إ. كولاى لا يصدق للفيل كما كان يظن عند بدء فك طلاسم هذه اللغة، إلا أننا نعلم كذلك أن هذين الكائنين على تباينهما حجما وتعقيدا وتطورا يتشاركان فى جوهر واحد هو لغة الحياة الصماء فى خلاياهما. فمن ألفباء من أربعة أحرف فقط تمثل وحدات البناء الكيميائية للدنا ينبثق بناء كامل من العمليات الحيوية يعتبر الإنسان أكثر تعبيراته تعقيدا. ويعد الكشف واستخدام تلك "الألفباء" فى تشكيل "كلمات وجمل" جديدة هو مصب اهتمام حقل البيولوجيا الجزيئية الحديثة وأعظم تحدياته.

والحياة موجودة من قبلنا بأربعة بليون سنة وستظل موجودة بعدنا (إن لم نقض عليها نحن بسياساتنا المدمرة)، وهى ليست بحاجة لمن يفك طلاسمها. ولكنها طوال فترة تواجدها على الأرض وهى تشق طريقها بنفسها وتعيد صياغة الكوكب لتوائم ظروفه وجودها عليه، سائرة فى ذلك فى خط مستقيم نحو زيادة حجم وقدرات ذلك العضو الحسى العصبى المسمى بالمخ والذى بلغ أكثر تعقيد له فى المخ البشرى، جاهلة كل شىء عن وجودها وجوهرها، حتى تسنى لهذا الأخير أن يحل اللغز ويقرأ كتابها.

وإنه ليحق لنا أفراد هذا النوع البشرى وأبناء هذا الجيل أن نفخر بأننا كنا أول من نجح فى قراءة سفر الحياة فى خلايا الكائنات الحية وفى خلايانا ذاتها، ويكفى أن القرن الماضى قد انتهى بإعلان مسودة الجينوم البشرى بعد خمسين عاما فقط من اكتشاف المادة الوراثية ومعرفة تركيبها، وهو ما سوف ينعكس بدوره على طبيعة وتوجهات البحوث البيولوجية فى القرن الحادى والعشرين. فمنذ التسعينيات والمجتمع

العلمى يعيش فى ثورة تشبه تلك التى واكبت اكتشاف الذرة فى مطلع القرن العشرين، وكأى ثورة علمية تكمن عظمتها فى مقدار ما تطرح من أسئلة جديدة أكثر من مقدار ما تمنح من إجابات. لقد أدت علوم الذرة وميكانيكا الكم إلى قلب الفيزياء الكلاسيكية رأسا على عقب، فما الذى سوف يؤدى إليه قراءة سفر الحياة ؟ وإذا كانت الحيطة والحذر يتوخيان بنا ألا ننجرف خلف نبوءات واسعة قد يثبت كذبها فيما بعد، إلا أننا نتنبأ بل ونؤمن أشد الإيمان أن أكثر الاكتشافات إثارة فى مجالات البيولوجيا والتى سوف يكشف عنها الغد هى تلك التى لا تزال تندرج اليوم فى خانة الأمور غير المتوقعة أو حتى المستحيلة.

لقد كان من حسن حظ البيولوجيا الجزيئية أن واكبت ثورة أخرى فى مجال تكنولوجيا المعلومات، فإن الحجم المذهل للبيانات الجزيئية وأنماطها الخفية والماكرة أدت إلى عدم وجود مناص من استخدام قواعد بيانات وأدوات تحليل تعمل بواسطة الكمبيوتر. فالتحدى الأكبر إذن هو إيجاد مداخل جديدة للتعامل مع حجم البيانات وتعقيدها ولمد الباحثين بطرق أسهل للوصول إلى المعلومة وأدوات الكمبيوتر حتى يتسنى لنا فهم أفضل لميراثنا الجينى ودوره فى كل من الصحة والمرض. وهو ما تصدى له علم المعلوماتية الحيوية. فالمعلوماتية الحيوية bioinformatics هى طريقة جديدة لقراءة البيولوجيا أكثر منها فرع من العلوم يستخدم الكمبيوتر فى الأبحاث البيولوجية كما يحلو للكثيرين النظر إليها (وهو ما يميزها عن الحوسبة الحيوية - biocomputing)، وهى تنظر للغة الحياة (تتابعات الدنا أو تتابعات البروتينات وأدبيات البيولوجيا) على أنها مجموعة من البيانات الخام data التى لا بد من معالجتها حتى يتم صياغة معنى لها، هذا المعنى هو المعلومة information التى هى غاية الباحث أى باحث. وتكون هذه المعالجة processing إما بتخزينها فى قواعد بيانات يسهل الوصول إليها، أو — وهذا هو الأهم — إيجاد الصلة بين هذه البيانات المختلفة أى ما هو التشابه بينها، وذلك فى ضوء محاور ثلاثة :

- يحدد تتابع الدنا تتابع البروتين
- يحدد تتابع البروتين تركيب البروتين (شكله الفراغى)
- يحدد تركيب البروتين وظيفة البروتين

وتجدر الإشارة إلى أن التشابه similarity فى حد ذاته هو مشكلة أنطولوجية، فحتى اليوم لا يمكن إيجاد مفهوم واحد للتشابه، فمثلا قد يتشابه نوعان من الأعمدة مثل الأعمدة الكورنثية والأعمدة الإيونية فى العمارة الإغريقية فى أن كلا منهما يتكون من بناء أسطوانى طويل ينتهى بإكليل مزخرف وهو تشابه تركيبى structural similarity يعكس انتماء كل منهما لنفس الفن، كما أنه قد يتشابه شيئان لا صلة تركيبية بينهما مثل معجون الأسنان وماكينة الحلاقة حيث يتشابهان فى أن كلا منهما يلعب دورا فى سيناريو الصباح اليومى وهو تشابه وظيفى functional similarity، كذلك قد يتشابه مقعد خشبى وثوب من الكتان فى أن كلا منهما قد صنع من مادة خام نباتية وهو ما يعرف بالتشابه التطورى evolutionary similarity (أى التشابه من حيث آليات التخليق). وفى البيولوجيا يتشابه الصرصور مع الذبابة تشابها تركيبيا (يتكون الجسم من راس وصدر وبطن ويحمل الصدر ستة أزواج من الأرجل وزوجين من الأجنحة بينما تحمل الراس العيون وقرنى الاستشعار وأجزاء الفم)، بينما تتشابه أجنحة الطيور مع أجنحة الخفافيش تشابها وظيفيا (وإن اختلفتا تركيبا)، بينما يتشابه الكنجارو مع الفأر مثلا تشابها تطوريا (فكل منهما حيوان ثديى وإن اختلفت آليات الولادة فى كل منهما تركيبا ووظيفة).

والخطوة الأولى لإيجاد اى تشابه بين البيانات الجزيئية تستخرج منه معلومة هى ترتيب التتابعات قيد الدراسة (سواء كانت تتابعات دنا او تتابعات بروتين أو تتابع دنا مع تتابع بروتين) بصورة يسهل معها تقدير التشابه ومعرفة طبيعته (تركيبى او وظيفى أو تطورى) وهو ما يعرف باسم رص التتابعات sequence alignment. وإذا كان كل من التركيب والوظيفة عادة ما يعكسان التطور فى البيولوجيا، حتى أنه ليصعب فصله عنهما او تفسيرهما بدونه، فإننا سوف نعمد بعد شرحنا لبعض اساسيات لغة الحياة إلى شرح التغيرات التطورية فى التتابعات وطرق تقدير معدلات وانماط تلك التغيرات وذلك فى معرض حديثنا عن حسابيات algorithms وبرمجيات software رص التتابعات والذى هو موضوع بحثنا هذا.

٦.٢. قواعد لغة الحياة.

٦.٢.١. تتابعات النيوكليوتيدات.

بخلاف بعض الفيروسات، فإن المادة الوراثية فى كل الكائنات الحية يحملها الحامض النووى الديؤكسى ريبوزى الذى يعرف اختصارا بالدنا DNA، والذى يتكون من شريطين يلتفان حول بعضهما البعض مشكلين لولبا مزدوجا يمينى الالتفاف. ويتكون كل شريط منهما من سلسلة من أربع نيوكليوتيدات nucleotides اثنان منهما من نوع البيورين ويرمز لهما بالحرف R (وهما الأدينين A والجوانين G) واثنان من البيريميدين ويرمز لهما بالحرف Y (وهما الثايمين T والسيتوزين C). ويرتبط شريطا الدنا من خلال روابط هيدروجينية بين نيوكليوتيداهما حيث دائما ما ترتبط نيوكليوتيدة من البيورين بنيوكليتيدة خاصة من البيريميدين، وتسمى الرابطة بين الأدينين والثايمين باسم الرابطة الضعيفة ويرمز لها بالحرف W، بينما تسمى الرابطة بين الجوانين والسيتوزين باسم الرابطة القوية ويرمز لها بالحرف S، وتكتب الروابط بوضع نقطتين بين رموز النيوكليوتيدات مثل C:G أو A:T فيما يعرف باسم **أزواج القواعد القانونية canonical base pairs**. ويوضح الجدول (٦-١) أبجدية كتابة تتابعات النيوكليوتيدات.

ويحدد اتجاه الروابط الفوسفودايسترية التى تربط بين نيوكليوتيدات الشريط الواحد فى السلسلة طبيعة الجزئ، وعليه فإن النيوكليوتيدات تكتب بترتيب النسخ أى فى الاتجاه 3′ ⟵ 5′ فقط، وبالنسبة لموقع معين على التتابع فإن كلا من الطرفين 5′ أو 3′ يسميان الجانب العلوى upstream أو الجانب السفلى downstream على التوالى. ويسمى الشريط الذى يحتوى على أكثر من ٥٠٪ من البيورينات باسم الشريط الثقيل heavy strand، بينما يسمى الشريط الذى يحتوى على أكثر من ٥٠٪ من البيريميدينات باسم الشريط الخفيف light strand.

جدول (٦-١) – أبجدية الدنا.

الحرف	التفسير
A	أدينين
C	سيتوزين
T	ثايمين
G	جوانين
W	روابط ضعيفة(A,T)
S	روابط قوية (C,G)
R	بيورينات (A,G)
Y	بيريميدينات (C,T)
K	كيتو (T,G)
M	أمينو (A,C)
B	C أو G أو T
D	A أو G أو T
H	A أو C أو T
V	A أو C أو G
N	A أو C أو T أو G
–	لا نيوكليوتيدة (فجوة)

أما الرنا RNA فإنه يتواجد إما فى صورة شريط مفرد أو مزدوج، وهو يختلف عن الدنا فى احتواءه على سكر ريبوز بدلا من سكر الديؤكسى ريبوز وفى أن نيوكليوتيدة اليوراسيل U تحل محل الثايمين، كما أن ازواج القواعد القانونية فيه تشمل G:U كذلك (بينما لا توجد رابطة G:T فى جزئ الدنا). وتسمى هذه النيوكليوتيدات مجتمعة (A، C، G، T، U) باسم النيوكليوتيدات القياسية standard nucleotides، ويلاحظ أن بعض جزيئات الرنا وخاصة الرنات الناقلة t-RNAs قد تحتوى على بعض النيوكليوتيدات غير القياسية والتى تشتق من النيوكليوتيدات القياسية. ويقاس طول الحمض النووى المفرد (سواء دنا أو رنا) بعدد نيوكليوتيداته، بينما يقاس الجزئ المزدوج بعدد أزواج القواعد (زق)(bp) base pairs.

وتسمى المادة الوراثية كلها فى الكائن الحى باسم الجينوم genome، وهو ينقسم إلى جزء يحمل جينات ويسمى الدنا الجينى genic DNA وجزء لا يحمل جينات يسمى الدنا غير الجينى nongenic DNA (وقد يكون هذا الجين حاملا لجينات لم يتم الكشف عنها بعد). ويسمى كل ما يتم نسخه من الجينوم من رنات باسم الترانسكريبتوم transcriptome، بينما يسمى كل ما يتم تشفيره (ترجمته) من بروتينات باسم البروتيوم proteome.

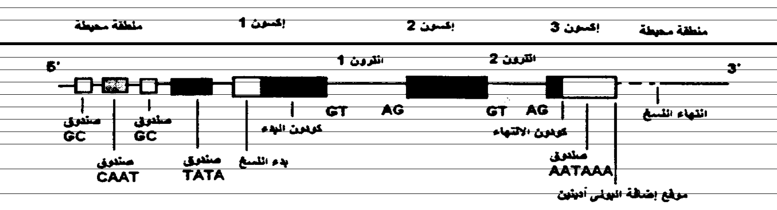
وبالرغم من أن تعريف الجين gene (وهو محور ابحاث علم الوراثة genetics عموما) لا يزال غامضا، إلا أن أكثر التعريفات الحديثة شيوعا هى تلك التى تعتبره قطعة من الدنا أو الرنا مسئولة عن وظيفة معينة، وليس بالضرورة أن يقتضى تحقيق هذه الوظيفة ترجمة أو حتى نسخا. وثمة ثلاثة أنواع من الجينات :

- جينات مشفرة للبروتينات protein-coding genes، وتلك يتم نسخها إلى رنات ومن ثم ترجمتها إلى بروتينات.
- جينات مخصصة للرنا RNA-specifying genes، وتلك يتم نسخها فقط.
- جينات لا يتم نسخها untranscribed genes.

ويسمى النوعان الأولان باسم الجينات التركيبية structural genes (وأحيانا تطلق هذه التسمية على النوع الأول فقط منها).

وتتكون الجينات المشفرة للبروتينات فى حقيقيات النواة (شكل ٦-١) من أجزاء يتم نسخها وأخرى لا يتم نسخها، والأجزاء التى لا يتم نسخها منطقتان محيطتان flanking regions، واحدة تقع عند الطرف '5 وتحتوى على تتابعات خاصة تسمى الإشارات signals تكون مسئولة عن حث عملية النسخ وضبط إيقاعها وتوقيتها وتعيين النسيج الذى سوف يتم التعبير عنها فيه، وأحيانا تسمى هذه التتابعات باسم المحفزات promoters، وتسمى المنطقة التى تحتويها باسم منطقة الحفز، وهى تحتوى على الإشارات التالية مثل صندوق TATA الذى يقع على بعد ١٩-٢٧ زوجا على الجانب العلوى من نقطة بدء النسخ ويتحكم فى اختيار هذه النقطة، وصندوق CAAT الذى يقع أكثر بعدا على الجانب العلوى، وصناديق CG المتباينة عددا وموقعا حول صندوق CAAT والتى تشترك

معه فى الارتباط ببوليميريز الرنا عند النسخ. وتجدر الإشارة إلى أن هذه التتابعات ليست بالضرورة موجودة فى كل الجينات. أما الجانب السفلى فيحتوى على إشارات إنهاء عملية النسخ. ونتيجة لقصر معلوماتنا عن كل من إشارات المنطقتين فإننا لا نزال إلى الوقت الحاضر عاجزين عن تحديد نقطتى بدء وانتهاء الجين بدقة.



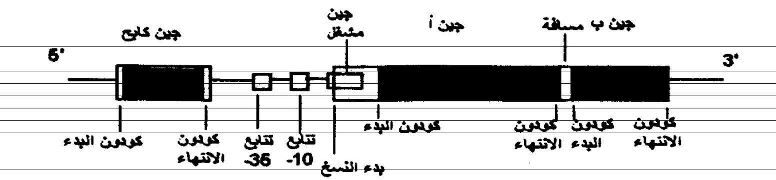
شكل (٦-١) — رسم تخطيطى لجين مشفر للبروتين نموذجى فى حقيقيات النواة ومكوناته التركيبية ويشير الخط المتقطع إلى احتمالية ألا يتماثل موقع إضافة البولى أدينين مع موقع انتهاء النسخ.

ويسمى الرنا المنسوخ داخل النواة باسم الرنا النووى غير المتجانس hnRNA، كما يسمى شريط الدنا الذى ينسخ منه الرنا باسم شريط الاتجاه المضاد antisense strand بخلاف الشريط المقابل الذى لا يتم نسخه ويكون متماثلا مع شريط الرنا والذى يسمى شريط الاتجاه الأصلى sense strand. ويبدأ النسخ عادة من الدنا عند نقطة تسمى موقع حث النسخ transcription initiation site (والتى تقابل قلنسوة cap الرنا)، بينما ينتهى عند نقطة تسمى موقع انتهاء النسخ transcription termination site (والتى ليس بالضرورة أن تقابل موقع إضافة ذيل البولى الأدينين على شريط الرنا الرسول mRNA الناضج فقد تقع بعد هذا الموقع). ويتكون بادئ الرنا الرسول pre-mRNA من إكسونات وانترونات، أما الانترونات introns فهى تلك التتابعات التى يتم قطعها والتخلص منها خلال عملية معالجة الرنا RNA processing، أما سائر التتابعات الأخرى والتى تبقى على طول شريط الرنا بعد عملية المعالجة ويتم الوصل بينها فتسمى إكسونات exons. وبالإضافة لعملية الوصل splicing يبلغ بادئ الرنا الرسول مرحلة

النضج بعد وضع قلنسوة من الجوانين الميثل عند الطرف '5 وتحلل النيوكليوتيدات التى قد تقع بعد نقطة إضافة ذيل عديد الأدينين ثم إضافة الأخير (والذى قد يبلغ طوله ١٠٠-٢٠٠ قاعدة ادينين). كذلك تتعرض جزيئات الرنا للعديد من عمليات التحكم النسخى وما بعد النسخى transcriptional and post-transcriptional processes.

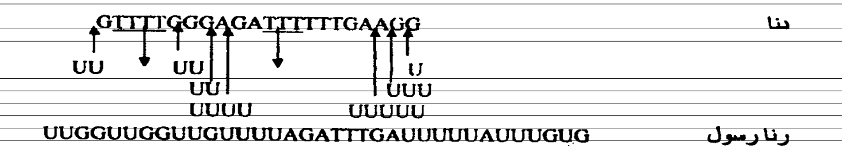
وتصنف الانترونات تبعا لطريقة الوصل إلى انترونات ذاتية الوصل self-splicing introns مثل انترونات جينات الميتوكوندريا والكلوروبلاست، وانترونات سبليسيوسومية spliceosomal introns مثل انترونات الجينات النووية والتى يتم انتزاعها من الرنا بواسطة معقد إنزيمى يدعى السبليسيوسوم spliceosome. وتسمى الأطراف '5 من الانترون بالمواقع المانحة donor sites وغالبا ما تكون GT، بينما تسمى الأطراف '3 بالمواقع المستقبلة acceptor sites وغالبا ما تكون AG فيما يعرف باسم قاعدة GT–AG. ويختلف عدد وتوزيع الانترونات من جين إلى آخر فى حقيقيات النواة، وإن كانت معظم جينات الفقاريات تتكون أساسا من انترونات

أما فى أوليات النواة، فإن جيناتها (شكل ٦-٢) لا تحتوى على انترونات، ولها محفزان للنسخ على الجانب العلوى يقعان على مسافة ١٠ قواعد (صندوق Pribnow ويتكون من التتابع TATAAT) و٣٥ قاعدة (تتابع TTGACA أو أحد تنويعاته) مع احتمال وجود محفزات أخرى. كذلك تنتظم الجينات التركيبية فى أوليات النواة على طول الجينوم مشكلة وحدة تعبير وراثى واحدة تنتج رنا رسولا طويلا يترجم إلى عدة بروتينات مختلفة، وتسمى هذه الجينات مجتمعة مع الجينات المنظمة لتعبيرها باسم الأوبيرون operon.



شكل (٦-٢) – رسم تخطيطى لأوبيرون مستحث فى أوليات النواة.

وسواء ترجم جزئ الرنا إلى بروتين أو لم يترجم فإنه عادة ما يعقب عملية النسخ عدد من التحويرات تعرف باسم تحرير الرنا RNA editing. وقد تبلغ عملية التحوير هذه حدا يصعب معه إيجاد التشابه بين جزئ الرنا وجزئ الدنا الذى نسخ منه، عندئذ يسمى قالب الجين باسم كريبتوجين (جين خفى) cryptogene (شكل ٦-٣).



شكل (٦-٣) – مقارنة بين منطقة ممثلة لشريط الاتجاه الصحيح لجين الميتوكوندريا المشفر للوحدة الثانوية III لأوكسيديز السيتوكروم c فى التريبانوزوما بروساى (*Trypanosoma brucei*) مع رنا رسول بعد التحرير. تشير الأسهم إلى أعلى إلى ادراج U بينما تشير الأسهم إلى أسفل إلى حذف T [من Feagin et al. (1988)].

كذلك يوجد عدد من الجينات لا تتم ترجمتها ولا حتى نسخها بيد أنها يكون لها وظائف حيوية حيث أنها تعمل كمواقع ارتباط للإنزيمات فى عمليات تضاعف الدنا (جينات التضاعف replicator genes) والعبور الوراثى (جينات العبور الوراثى recombinator genes) والتتابعات التيلوميرية telomeric sequences والانقسام الخلوى (جينات الانعزال segregator genes) وتتابعات مواقع الارتباط

attachment sites بالبروتينات الهيكلية والإنزيمات والهرمونات والنواتج الأيضية وتتابعات المواقع البنائية constructional sites والتى لها علاقة بشكل الكروموسومات.

وقد تظهر أيضا بعض قطع الدنا غير الجينية تشابها شديدا لجينات عاملة ولكنها تحتوى على عيوب (طفرات) تمنع التعبير عنها بصورة صحيحة وتسمى فى هذه الحالة جينات كاذبة pseudogenes. وعادة ما تكتب بوضع الرمز Ψ قبل اسم الجين العامل المشابه لها فى معظم الأدبيات وإن كانت فى قواعد بيانات الجينات الموجودة على الكمبيوتر يستبدل هذا الرمز بالحرف P.

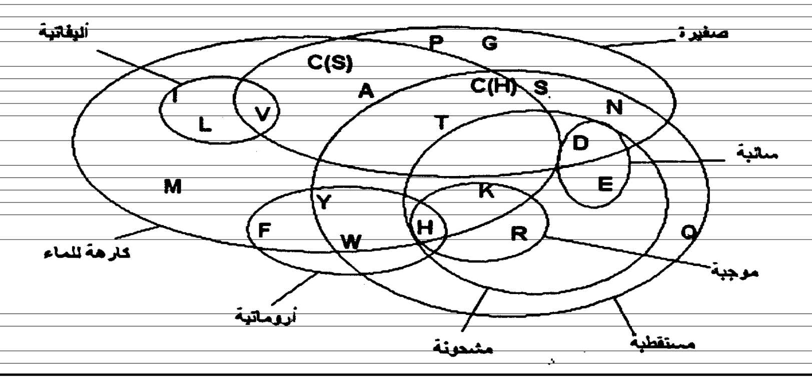
٦.٢.أ. تتابعات البروتينات.

تتكون البروتينات proteins فى كل الكائنات الحية من ٢٠ حمضا أمينيا amino acids أوليا مذكورة ابجديتها فى الجدول (٦-٢). ويتكون كل حمض أمينى من مجموعة أمين NH_2- ومجموعة كربوكسيل COOH- على جانبى ذرة كربون مركزية تسمى ألفا α carbon يرتبط بها كذلك ذرة هيدروجين ومجموعة السلسلة الجانبية R group-، وتكون الأخيرة هى المسئولة عن تمييز حمض أمينى عن الآخر، حيث أنها تتباين فى الحجم والشكل والشحنة والقدرة على تشكيل الروابط الهيدروجينية والتركيب والتفاعلية الكيميائية، ومن ثم فإن تصنيف البروتينات يتوقف على خواص سلاسلها الجانبية (شكل ٦-٤).

جدول (٦-٢) – ابجدية الأحماض الأمينية.

الحمض	الحرف	الحمض	الحرف
أرجينين	R	حمض الأسبرتك	D
أسبرجين	N	حمض لجلوتامك	E
ألانين	A	سستيين	C
أيزوليوسين	I	سيرين	S
برولين	P	فالين	V
تربتوفان	W	فينيل ألانين	F
تيروسين	Y	ليسين	K
ثريونين	T	ليوسين	L
جلوتامين	Q	ميثيونين	M
جليسين	G	هيستيدين	H

ويتكون البروتين من عدد من السلاسل البوليببتيدية لكل منها تتابعها الخاص من الأحماض الأمينية التى ترتبط مع بعضها البعض بواسطة روابط ببتيدية peptide bonds. ويسمى كل حمض أمينى فى السلسلة البوليببتيدية باسم فضالة residue، وتقرأ السلسلة من الطرف الأمينى أو النهاية النيتروجينية amino or N terminus إلى الطرف الكربوكسيلى carboxyl or C terminus. وكما سبق أن أشرنا فإن تتابع الأحماض الأمينية فى السلاسل البوليببتيدية للبروتينات هو الذى يحدد تركيب هذه البروتينات.



شكل (٦-٤) — رسم فن يبرز تقسيم الأحماض الأمينية الأولية العشرين إلى مجموعات متداخلة تبعا للحجم وتركيب السلسلة الجانبية والاستقطاب والشحنة والكراهية للماء. لاحظ وقوع السستيين فى مجموعتين كسستيين C(H) وكسستين C(S).

وثمة أربعة مستويات لتركيب البروتين، فالتركيب الأولى primary structure هو ببساطة الترتيب الخطى لفضالات الأحماض الأمينية على طول تتابع البوليببتيد. أما التركيب الثانوى secondary structure فيعنى الترتيب الفراغى أو التثنى folding لفضالات الأحماض الأمينية المتجاورة فى التركيب الأولى، ولعل أشهر أنواع هذا التركيب اثنين هما اللولب ألفا α helix وهو تركيب عصوى تنتظم فيه الأحماض الأمينية فى شكل حلزون يمينى الالتفاف تكون فيه سلاسلها الجانبية كلها للخارج وتنتظم الروابط الهيدروجينية بين مجاميع الأمينو ومجاميع الكربوكسيل على مسافة كل أربعة أحماض أمينية فى التتابع، وتركيب صفيحة بيتا β sheet وفيه تنتظم السلاسل المتوازية من الأحماض الأمينية بواسطة روابط هيدروجينية بين السلاسل المتجاورة. كذلك فبعض مناطق البروتين لا ينتظمها اى تركيب ثانوى ثابت وتظل فى التفاف عشوائى random coil. وعلى العموم فإن التركيب الثانوى لمعظم البروتينات يكون مجموعة من اللوالب ألفا والصفائح بيتا والالتفافات العشوائية.

ويسمى التركيب الفراغى ثلاثى الأبعاد للبروتين باسم **التركيب الثلاثى tertiary structure** ويعنى الترتيب الفراغى لفضالات الأحماض الأمينية غير المتجاورة فى التركيب الأولى، حيث ترتبط التراكيب الثانوية بواسطة قوى تساهمية وقوى غير تساهمية مثل الروابط الهيدروجينية والتفاعلات غير المحبة للماء وجسور الأملاح بين الفضالات مضادة الشحنة وروابط الديكبريتيد بين السستيينات. وفى هذا التركيب تعمد الأحماض الأمينية ذات السلاسل الجانبية الكارهة للماء إلى الاندفان فى قلب البروتين، بينما تنتظم الأحماض الأمينية ذات السلاسل الجانبية المحبة للماء على سطح البروتين.

وقد يتكون البروتين من أكثر من سلسلة بوليببتيدية واحدة تدعى كل منها فى هذه الحالة بوحدة ثانوية subunit، ويقال عن البروتين إن له **تركيبا رباعيا quaternary structure** وهو ما يعنى الترتيب الفراغى لهذه الوحدات الثانوية وطبيعة ارتباطها ببعضها البعض. وإلى اليوم، يعتبر التنبؤ بتركيب البروتين من تتابعه فقط أحد أكبر تحديات الكيمياء الحيوية والمعلوماتية الحيوية نتيجة لقصر معلوماتنا عن طبيعة تكوين الروابط.

وتحتاج بعض البروتينات إلى بعض المركبات غير البروتينية للقيام بوظيفتها، وتسمى تلك المواد باسم المجاميع الترقيعية prosthetic groups، ويطلق على معقد البروتين مع هذه المواد الترقيعية اسم **البروتين الكامل holoprotein**، بينما يسمى البروتين بدون هذه المواد باسم **البروتين الناقص apoprotein**.

٦.٢.٢. الشفرة الوراثية.

اثناء عملية تخليق البروتين تتم ترجمع تتابع الرنا الرسول إلى سلسلة بوليببتيدية بواسطة الرنات الناقلة tRNAs، حيث يكون لكل من الأحماض الأمينية العشرين رناه الناقل الخاص به والذى يحمل الكودون المضاد anticodon لثلاثى النيوكليوتيدات (الكودون codon) المحمول على الرنا الرسول. وتبدأ الترجمة عند كودون البدء initiation codon وتنتهى عند كودون الانتهاء termination codon، وتسمى المرحلة التى تصل إليها الترجمة على التتابع باسم **إطار القراءة reading frame**.

وما ان تنتهى الترجمة، حتى يتم تنشيط البروتين وظيفيا إثر بعض العمليات بعد النسخية مثل تحوير بعض الأحماض الأمينية الأولية أو نزع التتابعات الطرفية أو النقل داخل أو بين الخلايا أو إضافة المجاميع الترقيعية أو وصل البروتينات.

وكأى لغة فللغة الحياة قواعد تنظم العلاقة بين تتابعات الدنا والبروتينات تسمى الشفرة الوراثية genetic code (جدول ٦-٢). وبخلاف بعض الاستثناءات، فإن الشفرة الوراثية كونية universal (أى تشترك فيها كل الكائنات حقيقية وأولية النواة)، غير أن بعض العلماء يفضلون تسميتها بالشفرة الوراثية القياسية نتيجة لوجود هذه الاستثناءات. ويلاحظ أن بعض الأحماض الأمينية يشفر لها أكثر من كودون واحد فى ظاهرة تعرف باسم التنكسية degeneracy وفيها تسمى هذه الكودونات التى تشفر لنفس الحمض الأمينى باسم الكودونات المترادفة synonymous codons. كذلك فإن الكودونات المترادفة التى تختلف فقط فى النيوكليوتيدة الثالثة تسمى عائلة كودونية codon family.

وقد تغيب بعض الكودونات فى الجينات المشفرة للبروتينات تماما عن تتابع البروتين وتسمى كودونات غائبة absent codons، وأحيانا يعزى ذلك لغياب الرنا الناقل الخاص بها وتسمى الكودونات فى هذه الحالة باسم الكودونات غير المعينة unassigned codons ويمكن الكشف عنها بتوقف عملية الترجمة عند إطار قراءتها وعدم انفصال الرنا الرسول عن الريبوسوم.

٦.٢.٤. الطفرات.

تعرف الطفرات mutations بأنها أى تغير ينشأ فى تتابع الدنا نتيجة عادة للأخطاء التى تقع خلال التضاعف، وهى بذلك تعد المصدر الوحيد للاختلافات والمستحدثات فى التطور. حيث أن التطور evolution فى أبسط تعريفاته هو "الانحدار مع التحور" *"descent with modification"*، فإذا ما كانت هذه التحورات أو الاختلافات داخل نفس الأسرة (بين الآباء والأبناء) أو بين العشائر والأعراق داخل نفس النوع سمى

بالتطور الصغير microevolution، اما إذا تعدت هذه الاختلافات مستوى النوع وأدت إلى نشوء انواع جديدة (وذلك بوضع حواجز تناسلية بين الأفراد الطافرة) فإنه يسمى بالتطور الكبير macroevolution. ومن الناحية التطورية تعتبر

جدول (٦-٣) — الشفرة الوراثية الكونية (وتظهر الاختلافات الموجودة فى الشفرة الوراثية لميتوكوندريا الفقاريات بين قوسين).

الكودون	الحمض الأميني	الكودون	الحمض الأميني	الكودون	الحمض الأميني	الكودون	الحمض الأميني
UUU	F	UCU	S	UAU	W	UGU	C
UUC	F	UCC	S	UAC	W	UGC	C
UUA	L	UCA	S	UAA	Stop	UGA	Stop (W)
UUG	L	UCG	S	UAG	Stop	UGG	W
CUU	L	CCU	P	CAU	H	CGU	R
CUC	L	CCC	P	CAC	H	CGC	R
CUA	L	CCA	P	CAA	Q	CGA	R
CUG	L	CCG	P	CAG	Q	CGG	R
AUU	I	ACU	T	AAU	N	AGU	S
AUC	I	ACC	T	AAC	N	AGC	S
AUA	I (M)	ACA	T	AAA	K	AGA	R (Stop)
AUG	M	ACG	T	AAG	K	AGG	R (Stop)
GUU	V	GCU	A	GAU	N	GGU	G
GUC	V	GCC	A	GAC	N	GGC	G
GUA	V	GCA	A	GAA	E	GGA	G
GUG	V	GCG	A	GAG	E	GGG	G

الطفرات التى تحدث فى الأنسجة التناسلية هى الأكثر اهمية حيث انها تلك التى يتم توريثها للأجيال التالية، غير انه من الناحية الطبية فإن للطفرات الجسمية اهمية كبرى كذلك إذ انها قد تكون مسئولة عن بعض الأمراض مثل السرطان مثلا.

وتنقسم الطفرات إلى :

١. طفرات إحلالية substitution mutations : وفيها يتم إحلال نيوكليوتيدة محل اخرى، فإذا حلت نيوكليوتيدة بيورينية محل نيوكليوتيدة اخرى من نفس النوع أو حلت نيوكليوتيدة بيريميدينة محل نيوكليوتيدة اخرى من نفس النوع سمى الإحلال transition (وهو اربعة انواع : A → G, G → A, C → T, T → C)، اما إذا حلت نيوكليوتيدة بيورينية أو بيريميدينة محل نيوكليوتيدة اخرى من النوع الآخر سمى

الإحلال transversion (وهو ثمانية أنواع : → A → C, A → T, C → A, C (G, T → A, T → G, G → C, G → T). وقد تترجم هذه الطفرات على مستوى تتابع البروتين فإذا لم تحدث تغيرا سمين طفرات مترادفة synonymous، أما إذا أحدثت فإنها تسمى طفرات غير مترادفة nonsynonymous. وفي هذه الحالة إذا تم استبدال حمض أميني بآخر على تتابع البروتين فإنه يسمى استبدالا replacement. ويجدر الإشارة إلى أنه ليست كل الطفرات المترادفة تكون صامتة silent أي لا تغير من البروتين المنتج، وإنما قد يؤثر بعضها على عملية الترجمة كتحويل إكسون إلى انترون وما إلى ذلك. كذلك تصنف الطفرات غير المترادفة إلى طفرات خاطئة missense تغير حمضا امينيا بآخر، وأخرى لا معنى لها nonsense وهي تلك التي تحول كودونا مشفرا لحمض أميني إلى كودون انتهاء الترجمة ومن ثم ينتج البروتين ناقصا.

٢. طفرات إعادة التوليفات الوراثية recombinations : وتحدث نتيجة للعبور الوراثي crossing over أو نتيجة للتحور الجيني gene conversion.

٣. طفرات الحذف والإدراج deletions and insertions (indels) : وتحدث نتيجة العبور الوراثي غير المتساوي unequal crossing over والانتقال الوراثي DNA transposition والانتقال الأفقي للجينات horizontal gene transfer، وقد تؤدي هذه الطفرات إلى تغيير إطار القراءة ومن ثم تركيب البروتين كله، عندئذ تسمى طفرات نقل الإطار frameshift mutations.

٤. طفرات الانقلاب inversions، وتحدث نتيجة لكسور في الكروموسومات.

وتجدر الإشارة إلى أن الطفرات تختلف في معدلاتها بين الأنواع، فمثلا يبلغ معدل الطفور في فيروس الإنفلونزا ٢ مليون مرة أكثر من معدل الطفور في الدنا النووي للفقاريات. كما تختلف الطفرات في توزيعها على طول الجينوم حيث تعرف المناطق ذات المعدلات العالية بالنقاط الساخنة hotspots مثل الجزر الغنية بالستيوزين والجوانين CpG isochors. كذلك تختلف في أنماطها فنجد أن الـtransitions أكثر شيوعا من

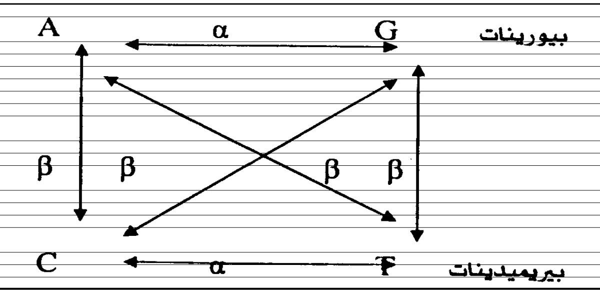
الـtransversions، كذلك فإن النيوكليتيدتين C وG أكثر تطفرا من النيوكليوتيدتين A وT.

ومن المعروف ان الطفرات تحدث "عشوائيا"، ولا يعنى هذا أن عشوائية الطفرات ترتبط بموقع الطفرة على الجينوم أو جميع أنواع الطفرات لها نفس التكرار، وإنما يرتبط مفهوم عشوائية الطفرة بتأثيرها على مواءمة الفرد حاملها. أى أن لأى طفرة نفس احتمال الحدوث سواء كانت مفيدة أو ضارة، وهو ما عده دبزانسكى "نقصا فى الطبيعة" (Dobzhansky, 1970).

٦.٣. التغيرات التطورية فى التتابعات.

٦.٣.١. نموذج كيميورا ذو المقياسين للتغير فى تتابع النيوكليوتيدات.

لا يولد تتابع من عدم وإنما يعزى أى اختلاف بين تتابعين أو أكثر نتيجة للطفرات التى حدثت لكل منهما منذ زمن انحدارهما عن تتابعهما السلفى المشترك. ولدراسة ديناميات الإحلالات النيوكليوتيدية فلا بد من وضع عدة فرضيات بالنسبة لاحتمال إحلال نيوكليوتيدة محل أخرى. ولقد تم اقتراح عدة نماذج رياضية (راء: Li, 1997)، ولكننا سوف نقصر هنا مناقشتنا على أحد أبسط هذه النماذج وهو النموذج ذى المقياسين لكيميورا Kimura's two-parameter model (1980)، والذى يقوم على افتراض أن معدل الإحلالات من نوع الـ transition ويرمز له بالرمز α لا يتساوى مع معدل الإحلالات من نوع الـtransversion ويرمز له بالرمز β (شكل ٦-٥).



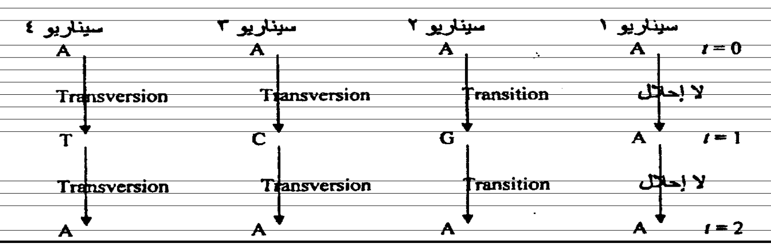
شكل (٦-٥) – نموذج الإحلال النيوكليوتيدى ذى المقياسين، حيث قد لا يتساوى معدل الإحلالات من نوع الـtransition ورمزه α مع معدل الإحلالات من نوع الـ transversion ورمزه β.

وعليه فإنه إذا ما كان لدينا موقع يحمل النيوكليوتيدة A عند الزمن $t = 0$، فإنه بعد وحدة زمنية واحدة يكون احتمال ان تتغير A إلى G هو α واحتمال أن تتغير إلى C أو T هو 2β، وهكذا يكون احتمال أن يظل هذا الموقع يحمل النيوكليوتيدة A بعد وحدة زمنية واحدة هو

$$P_{AA(1)} = 1 - \alpha - 2\beta \quad \text{...............(١)}$$

اما عند $t = 2$، فثمة أربعة احتمالات : (١) ألا تتغير A فى الوحدتين الزمنيتين، (٢) ان تتغير A إلى G فى $t = 1$ ثم ترجع G إلى A فى $t = 2$ بالـtransition، (٣) ان تتغير A إلى C فى $t = 1$ ثم ترجع C إلى A فى $t = 2$ بالـtransversion، (٤) ان تتغير A إلى T فى $t = 1$ ثم ترجع T إلى A فى $t = 2$ بالـtransversion (شكل ٦-٦)، وعليه فإن

$$P_{AA(2)} = (1 - \alpha - 2\beta)P_{AA(1)} + \beta P_{TA(1)} + \beta P_{CA(1)} + \alpha P_{GA(1)} \quad \text{... (٢)}$$



شكل (٦-٦) – اربعة سيناريوهات للحصول على A عند t = 2 تبعا لنموذج كيميورا ذى المقياسين إذا كانت A موجودة على نفس الموقع عند t = 0.

وهو ما يمكن تعميمه إلى:

$$P_{AA(t+1)} = (1 - \alpha - 2\beta)P_{AA(t)} + \beta P_{TA(t)} + \beta P_{CA(t)} + \alpha P_{GA(t)} \quad \text{..(٣)}$$

وبالتفاضل فى التغير الزمنى

$$\delta P_{AA(t)}/\delta t = -(\alpha + 2\beta)P_{AA(t)} + \beta P_{TA(t)} + \beta P_{CA(t)} + \alpha P_{GA(t)} \quad \text{.(٤)}$$

وبالمثل يمكننا الحصول على معادلات احتمالات الإحلالات الثلاثة، والتى منها نصل إلى

$$P_{AA(t)} = ¼ + ¼\ e^{-4\beta t} + ½\ e^{-2(\alpha+\beta)t} \quad (٥)$$

أى أنه عند الاتزان ($t = \infty$)، يكون احتمال أى عدم إحلال $X_{(t)}$ يساوى ¼.

وبعامة فإنه عند الزمن t، يكون احتمال الإحلال من نوع الـ transition

$$Y_{(t)} = ¼ + ¼\ e^{-4\beta t} - ½\ e^{-2(\alpha+\beta)t} \quad (٦)$$

بينما يكون احتمال الإحلال من نوع الـ transversion

$$Z_{(t)} = ¼ - ¼\ e^{-4\beta t} \quad (٧)$$

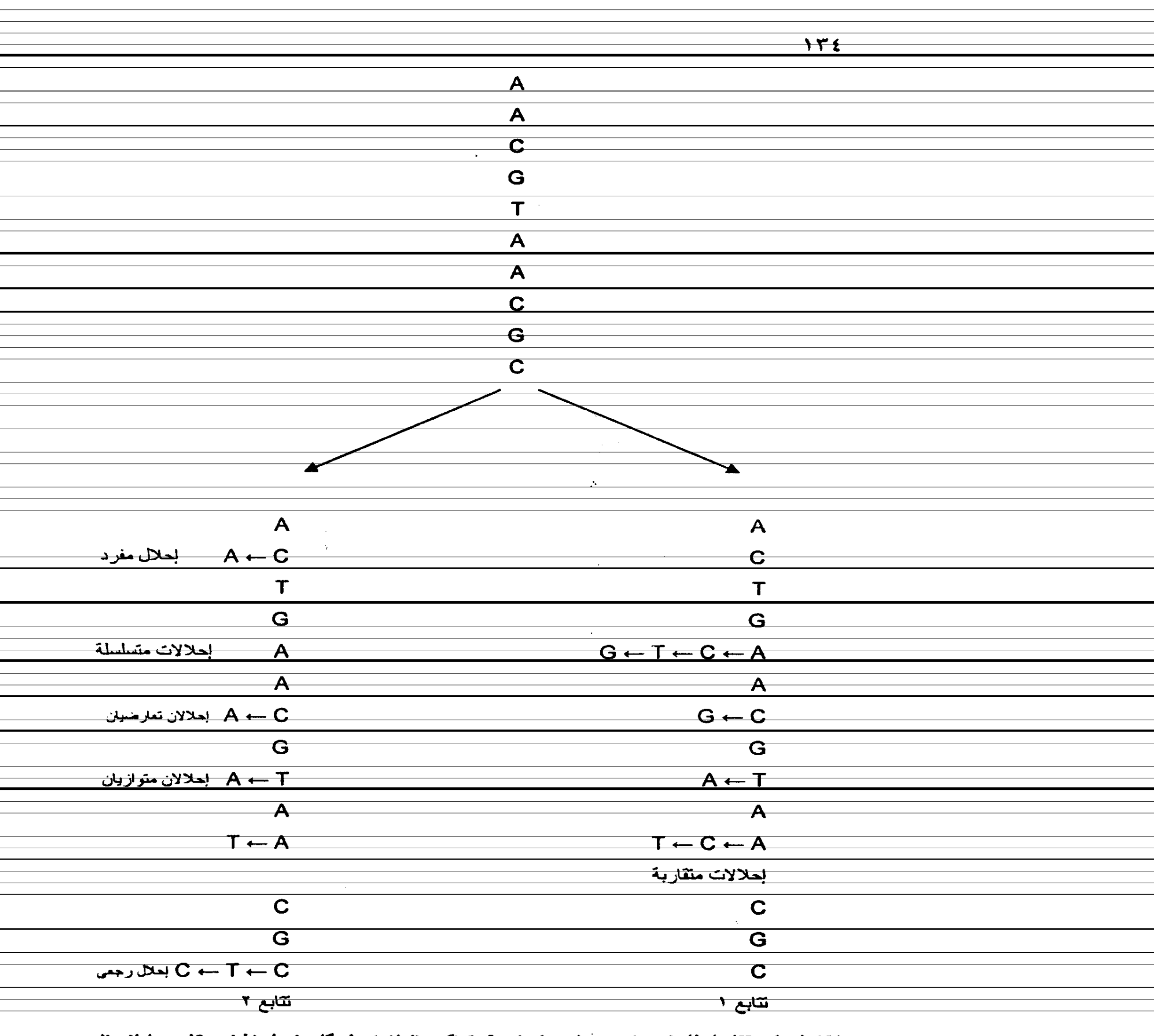
حيث أن لكل نيوكليوتيدة نوع واحد من الإحلالات من نوع الـ transition ونوعين من نوع الـtransversion K، وبالتالى فإن

$$X_{(t)} + Y_{(t)} + 2Z_{(t)} = 1$$

٦.٣.٢. عدد الإحلالات النيوكليوتيدية بين تتابعين للدنا.

يعتبر عدد الإحلالات النيوكليوتيدية بين تتابعين هو أصل وأكثر المقاييس شيوعا لتقدير مدى الاختلاف بينهما. فإذا اختلف تتابعان طول كل منهما N فى عدد n من المواقع، فإن النسبة n/N تعرف باسم درجة التشعب degree of divergence بينهما أو مسافة هامنج Hamming distance، وعادة ما تكتب كنسبة مئوية. ويعيب هذه الطريقة عدم قدرتها على الكشف عن الإحلالات المتعددة multiple substitutions or multiple hits فى نفس الموقع، كأن تتغير A إلى C ثم إلى T فى تتابع وتتغير إلى T مباشرة فى التتابع الآخر فلا يكشف عنها بالرغم من وجود ثلاثة إحلالات بين هذين التتابعين (شكل ٦-٧).

وهكذا يفضل التعبير عن مقدار هذا الاختلاف باستخدام عدد الإحلالات النيوكليوتيدية لكل موقع نيوكليوتيدى number of nucleotide substitutions per nucleotide site ويرمز لها بالرمز K بدلا من عدد الإحلالات النيوكليوتيدية الكلية بين التتابعين، خاصة إذا اختلف التتابعان فى أطوالهما. وتختلف طريقة حسابها بين إذا ما كان التتابع مشفرا أو غير مشفر للبروتين نتيجة لاختلاف معدلات الإحلال بينهما.



شكل (٦-٧) – تتابعا دنا منحدران من سلف مشترك وقد تراكمت الطفرات فى كل منهما منذ زمن تشعبهما. لاحظ حدوث ١٢ طفرة بعدة أنواع من الإحلالات حتى أن ٢ منها فقط هى التى يمكن مشاهدتها.

ففى حالة التتابعات غير المشفرة للبروتينات، وبافتراض أن عدد المواقع المقارنة بين التتابعين يساوى L ونسبة الاختلافات من نوعى الـtransition والـtransversion يساويان P و Q على التوالى، فإن K تساوى

$$K = ½ \ln(1 / 1 – 2P – Q) + ¼ \ln(1 / 1 – 2Q) \quad .... (٨)$$

وذلك بتباين معاينة يساوى

$$V(K) = 1/L [P(1 / 1 – 2P – Q)^2 + Q(1 / 2 – 4P – 2Q + 1 / 2 – 4Q)^2 – ((P / 1 – 2P – Q) + (Q / 2 – 4P – 2Q) + (Q / 2 – 4Q))^2] \quad .... (٩)$$

ويلاحظ أننا قد استخدمنا هنا نموذج كيميورا ذا القياسين، وأن هذه المعادلات سوف تتغير إذا ما كنا قد استخدمنا نماذج ذات مقاييس أخرى.

أما بالنسبة للتتابعات المشفرة للبروتينات، فتعتبر عملية تحديد عدد الإحلالات النيوكليوتيدية أكثر تعقيداً من حالة التتابعات غير المشفرة للبروتينات حيث يجب التمييز بين الإحلالات المترادفة والإحلالات غير المترادفة.

وعند مقارنة تتابعين مشفرين للبروتين يتم إقصاء كودونى البدء والانتهاء حيث أن هذين الكودونين نادرا ما يتغيران مع الوقت.

أما عند حساب عدد الإحلالات لكل موقع لكل من الإحلالات المترادفة وغير المترادفة على حدة فلا بد أولا من إيجاد القاسمين المناسبين، أى عدد المواقع المترادفة وغير المترادفة على التوالى. وهو ما يصعب عمله لسببين :

(١) يتغير تصنيف أى موقع مع الزمن، فمثلا يعتبر الموقع الثالث للكودون CGG المشفر للأرجينين مترادفا، ولكن إذا تغير الموقع الأول إلى T فإن الموقع الثالث والذى يصبح للكودون TGG المشفر للتربتوفان يصير غير مترادف.

(٢) بعض المواقع ليست مترادفة تماما أو غير مترادفة تماما، فمثلا إذا حدث إحلال من نوع الـtransition فى الموقع الثالث للكودون GAT المشفر لحمض الأسبرتيك فإنه سوف يكون مترادفا، أما إذا حدث إحلال من نوع الـ transversion فإنه سوف يكون غير مترادف، مع الأخذ فى الاعتبار أن الإحلالات من نوع الـtransition أكثر تكرارا من نوع الـtransversion.

وفى سبيل ذلك تم وضع عدة نماذج سوف نتناول أحدها وهو النموذج الذى وضعه Li *et al.* (1985) حيث أنه الأكثر شيوعا، وفيه يتم تقسيم المواقع النيوكليوتيدية إلى ثلاث فئات :

- موقع عديم التنكسية nondegenerate، إذا ما كانت جميع التغيرات عند هذا الموقع غير مترادفة، مثل الموقعين الأولين للكودون TTT المشفر للفينيل ألانين.
- موقع ثنائى التنكسية twofold degenerate، إذا ما كان أحد التغيرات الثلاث عند هذا الموقع مترادفا، مثل الموقع الثالث للكودون TTT.
- موقع رباعى التنكسية fourfold degenerate، إذا ما كانت جميع التغيرات عند هذا الموقع مترادفة، مثل الموقع الثالث من الكودون GTT المشفر للفالين.

وعليه فإنه عند مقارنة تتابعين واتباع القواعد السابقة فإنه يتم عد أنواع المواقع الثلاثة لكل تتابع ثم حساب متوسط كل منها حيث يرمز للمواقع عديمة التنكسية بالرمز L_0، وللمواقع ثنائية التنكسية L_2، وللمواقع رباعية التنكسية L_4. ثم يتم حساب عدد الإحلالات لكل نوع من المواقع على حدة، وذلك بعد تصنيف الاختلافات النيوكليوتيدية داخل كل فئة إلى اختلافات من نوع الـ transition ويرمز لها S_i واختلافات من نوع الـ transversion ويرمز لها بالرمز V_i حيث $i = 0, 2$ or 4 تبعا لدرجة التنكسية. ويلاحظ أنه بالتعريف فإن أى إحلال فى موقع عديم التنكسية يكون غير مترادف، بينما أى إحلال فى موقع رباعى التنكسية يكون مترادفا، ومن هنا تكمن المشكلة فى المواقع ثنائية التنكسية حيث تكون جميع الإحلالات من نوع الـ transition مترادفة، بينما تكون جميع الإحلالات الأخرى من نوع الـtransversion غير مترادفة. كذلك لا بد من الانتباه إلى الشفرة الوراثية المستخدمة، حيث قد تختلف تنكسية الكودون من الشفرة الكونية إلى شفرة ميتوكوندريا الفقاريات مثلا، حيث أنه فى الأخيرة لا توجد

استثناءات، بينما يوجد استثناءان فى الشفرة الكونية وهما الموقع الأول للكودونات الأربعة المشفرة للأرجينين (CGA, CGG, AGA, AGG) والموقع الثالث للكودونات الثلاثة المشفرة للأيزوليوسين (ATT, ATC, ATA).

وتحسب نسبة الاختلافات من نوع الـtransition عند مواقع ذات تنكسية من الدرجة i بين تتابعين من المعادلة

$$P_i = S_i / L_i \quad ....(١٠)$$

وبالمثل، تحسب نسبة الاختلافات من نوع الـtransversion عند مواقع ذات تنكسية من الدرجة i بين تتابعين من المعادلة

$$Q_i = V_i / L_i \quad ....(١١)$$

وباستخدام نموذج كيميورا ذى المقياسين لتقدير عدد الإحلالات من نوعى الـ transition ويرمز لها بالرمز A_i والـtransversion ويرمز لها بالرمز B_i من المعادلتين

$$A_i = ½ \ln(a_i) – ¼ \ln(b_i) \quad ....(١٢)$$

و

$$B_i = ½ \ln(b_i) \quad ....(١٣)$$

وذلك بتباين

$$V(A_i) = [a_i^2P_i + c_i^2Q_i – (a_iP_i + c_iQ_i)^2] / L_i \quad ....(١٤)$$

و

$$V(B_i) = b_i^2Q_i(1 – Q_i) / L_i \quad ....(١٥)$$

حيث L_i هو عدد المواقع ذات التنكسية من الدرجة i، و$a_i = 1 / (1 – 2P_i – Q_i)$، و$b_i = 1 / (1 – 2Q_i)$، و$c_i = (a_i – b_i) / 2$. ويكون العدد الكلى للإحلالات لكل موقع ذى تنكسية من الدرجة i ويرمز له بالرمز K_i وهو

$$K_i = A_i + B_i \quad ....(١٦)$$

بتباين

$$V(K_i) = [a_i^2P_i + d_i^2Q_i - (a_iP_i + d_iQ_i)^2] / L_i \quad (١٧)$$

حيث $d_i = b_i + c_i$.

ومما سبق وبعد التصحيح الذى وضعه Pamilo and Bianchi و Li (1993) (1993) يكون عدد الإحلالات المترادفة لكل موقع مترادف number of synonymous substitutions per synonymous site ورمزه K_s يساوى

$$K_s = [L_2A_2 + L_4A_4 / L_2 + L_4] + B_4 \quad (١٨)$$

بتباين

$$V(K_s) = [L_2^2V(A_2) + L_4^2V(A_4) / L_2 + L_4] + V(B_4) - [2b_4Q_4(a_4P_4 - c_4(1-Q_4)) / L_2 + L_4] \quad (١٩)$$

بينما يكون عدد الإحلالات غير المترادفة لكل موقع غير مترادف number of nonsynonymous substitutions per nonsynonymous site ورمزه K_A يساوى

$$K_A = A_0 + [L_0B_0 + L_2B_2 / L_0 + L_2] \quad (٢٠)$$

وتباينه

$$V(K_A) = V(A_0) + [L_0^2V(B_0) + L_2^2V(B_2) / (L_0 + L_2)^2] - [2b_0Q_0(a_0P_0 - c_0(1-Q_0)) / L_0 + L_2] \quad (٢١)$$

٦.٣.٢. عدد إبدالات الأحماض الأمينية بين بروتينين.

يمكن حساب النسبة المشاهدة للأحماض الأمينية المختلفة بين تتابعين من الأحماض الأمينية عند مقارنتهما على النحو التالى

$$p = n / L \quad (٢٢)$$

حيث n هى عدد الأحماض الأمينية بين التتابعين و L هى طول التتابعين المرصوصين.

وباستخدام توزيع بواسون يمكن تحويل p إلى عدد إبدالات الأحماض الأمينية لكل موقع على النحو التالى

$$d = -\ln(1 - p) \quad (٢٣)$$

وتباينه

$$V(d) = p / L(1 - p) \quad (٢٤)$$

٦. ٤ . رص التتابعات فى أزواج.

٦. ٤. ١. أساسيات رص التتابعات.

تتضمن مقارنة تتابعين متماثلين تعيين مواضع الحذف والإدراج الذين يمكن أن يكونا قد وقعا فى أى من التتابعين منذ تشعبهما عن سلفهما المشترك، وهو ما يطلق عليه عملية رص التتابعات sequence alignment. وسوف نشرح هنا هذه العملية مستخدمين تتابعات الدنا مع العلم بأن عملية رص تتابعات الأحماض الأمينية تخضع لنفس مبادئ وخطوات رص تتابعات الدنا. بل فى الحقيقة تعتبر النتائج المتحصل عليها من رص تتابعات الأحماض الأمينية أجدر بالثقة من نتائج رص تتابعات الدنا لسببين رئيسيين : (١) تتغير الأحماض الأمينية بصورة أقل من النيوكليوتيدات خلال التطور، (٢) هناك ٢٠ حمضا أمينيا بينما لا يتعدى عدد النيوكليوتيدات أكثر من ٤ ومن ثم فإن احتمال تماثل موقعين نتيجة للصدفة يكون أقل على مستوى الأحماض الأمينية عنه على مستوى النيوكليوتيدات

ويتكون رص تتابع دنا من سلسلة من ازواج القواعد (قاعدة من كل تتابع)، ومن ثم يكون ثمة ثلاثة أنواع من أزواج القواعد :

- توافقات matches، وفيها تظهر نفس النيوكليوتيدة فى نفس الموقع فى التتابعين ومن ثم نفترض عندئذ أن هذه النيوكليوتيدة لم تتغير منذ تشعب التتابعين من سلفهما المشترك.
- عدم توافقات mismatches، وفيها يكون فى كل تتابع نيوكليوتيدة مختلفة فى نفس الموقع، ومن ثم فإن إحلالا واحدا على الأقل قد وقع منذ تشعبهما من سلفهما المشترك.

- فجوات gaps، وتتكون من قاعدة موجودة فى أحد التتابعين تقابلها قاعدة معدومة null base فى نفس الموقع على التتابع الآخر، ويرمز للقواعد المعدومة بالرمز –، وتشير الفجوة إلى حدوث حذف فى أحد التتابعين أو إدراج فى الآخر. وعلى العموم لا ينبؤنا الرص فى ذاته بأى منهما هو الذى قد وقع بالفعل.

وباعتبار وجود تتابعين من الدنا هما A وB طول كل منهما هو m و n على التوالى، فإذا رمزنا إلى عدد الأزواج المتوافقة بالرمز x وعدد الأزواج غير المتوافقة بالرمز y وعدد الفجوات بالرمز z فإن

$$n + m = 2(x + y) + z \quad \text{.............. (٢٥)}$$

وعادة ما يتم التمييز بين الفجوات الطرفية terminal gaps والفجوات الداخلية internal gaps. فمثلا يتم إقصاء الفجوات الطرفية من الحسابات عند مقارنة جزء من التتابع مع تتابع كامل، أو يتم إقصاء الفجوات الداخلية عند رص تتابع جينومى مع تتابع رنا رسول حيث ان الأول يكون محتويا على انترونات وهو ما يفتقر إليه تتابع الرنا الرسول.

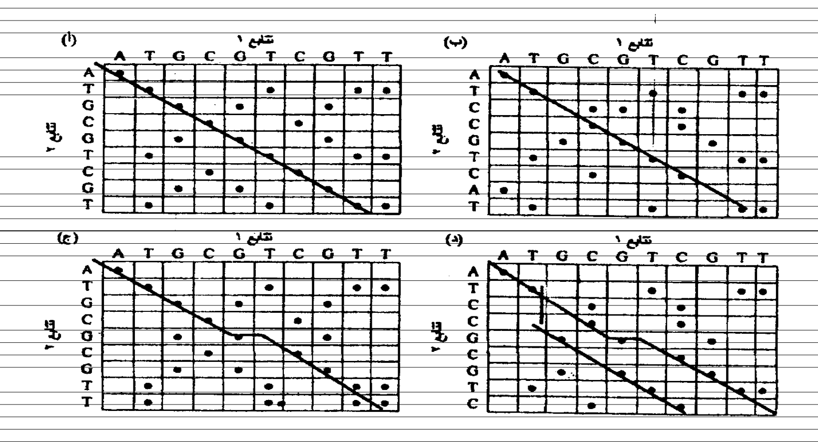
ومن حيث المبدأ ومن زاوية تطورية، فإن كل زوج فى الرص يمثل تماثلا موضعيا positional homology، أى ان عضوى الزوج منحدران من نيوكليوتيدة سلفية مشتركة. ومن ثم فإن أى خطأ فى الرص يعنى بالضرورة غياب معنوية الرص حيث أننا نعجز حينئذ عن تقدير عدد الإحلالات بصورة صحيحة.

فإذا علمنا أن الرص هو الخطوة الأولى فى العديد من الدراسات الجينومية وأن أى خطأ فى الرص سوف يتضاعف فى العمليات الحسابية اللاحقة، فإنه يتحتم علينا أن نقوم بعملية الرص بمنتهى الدقة. وعلى المرء أن يتخلص من كل الأجزاء المثيرة للشك من الرصيص قبل أن يبدأ فى التحليل، حتى وإن أدى ذلك إلى نقص طول الرص بصورة ملحوظة وما يصحبه من ارتفاع فى قيمة خطأ المعاينة عند تقدير عدد الإحلالات النيوكليوتيدية بين التتابعين.

٢.٣. رص التتابعات باستخدام طريقة مصفوفة النقاط.

لعل طريقة مصفوفة النقاط dot matrix method التى وضعها Gibbs and McIntyre (1970) هى اشهر واسهل طرق رص التتابعات، وفيها تتم كتابة التتابعين قيد الدراسة كعناوين راسية وافقية لأعمدة وصفوف مصفوفة ذات بعدين (شكل ٦-٨)، ثم توضع نقطة على رسم مصفوفة النقاط dot matrix plot عند موضع تماثل النيوكليوتيدات فى التتابعين. ومن ثم فإن نقطة عند (x,y) تعنى أن نيوكليوتيدة عند الموقع x فى التتابع الأول هى نفسها النيوكليوتيدة عند الموقع y فى التتابع الثانى.

ويعرف الرصيص alignment بأنه طريق فى المصفوفة يبدأ من أعلى العناصر إلى اليسار وينتهى عند أسفل العناصر إلى اليمين. ومن ثم فإن هناك أربع خطوات لهذا الطريق : (١) خطوة قطرية عبر نقطة تعنى توافقا، (٢)خطوة



شكل (٦-٨) – مصفوفات نقاط لرص تتابعين نيوكليوتيديين. (أ) التتابعان متماثلان فى الجزء المرصوص. (ب) يختلف التتابعان عن بعضهما البعض ولكن لا يحتوى الجزء المرصوص على فجوات. (ج) يحتوى الرصيص على فجوة، ولكن فيما عدا ذلك يتماثل التتابعان. (د) يظهر رصيصان يحتوى كل منهما على فجوات وعدم توافقات مما يدفع إلى استخدام بعض الحسابات للاختيار بينهما.

قطرية عبر عنصر خال فى المصفوفة تعنى عدم توافق، (٣) خطوة أفقية تعنى وجود نيوكليوتيدة معدومة (فجوة) فى التتابع على رأس المصفوفة، (٤) خطوة رأسية تعنى وجود نيوكليوتيدة معدومة (فجوة) فى التتابع على جانب المصفوفة.

فإذا كان التتابعان تامى التماثل (أو إذا تمت مقارنة تتابع بنفسه)، فسيتم وضع نقاط فى كل العناصر القطرية فى المصفوفة (شكل ٦-٨أ). أما إذا اختلف التتابعان عن بعضهما البعض فقط عن طريق إحلالات فسيتم وضع نقاط على معظمة العناصر القطرية فى المصفوفة (شكل ٦-٨ب). أما إذا حدث إدراج فى أحد التتابعين بلا أى إحلال فسوف توجد منطقة داخل المصفوفة يتم فيها نقل قطر الرصيص إما رأسيا أو أفقيا (شكل ٦-٨ج). وفى كل من الحالات السابقة الثلاث، كان الرصيص يشير إلى نفسه، أما إذا اختلف التتابعان نتيجة لوجود فجوات وإحلالات بينهما، فإنه يصعب تعيين موقع الفجوات ومن ثم يتحتم علينا الاختيار من بين أكثر من رصيص (شكل ٦-٨د)، وفى مثل هذه الحالات يفضل استخدام طرق أكثر ثقة من طريقة مصفوفة النقاط.

ويلاحظ فى الشكل (٦-٩أ) أنه فى معظم الحالات تكون مصفوفة النقاط شديدة التشوش أى يتم شغر عناصر أخرى فى المصفوفة غير تلك الممثلة للرصيص الحقيقى بنقاط تغطى على الرصيص، حيث أنه عند مقارنة تتابعين من الدنا، فإن حوالى ٢٥٪ من عناصر المصفوفة يتم شغرها بالصدفة وحدها.

وثمة مقياسان لتحديد عدد التوافقات الزائفة ومن ثم درجة وضوح رسم مصفوفة النقاط وهما حجم النافذة window size والشرطية stringency. وعليه فإنه بدلا من استخدام مواقع نيوكليوتيدية مفردة، فإنه يمكن مقارنة التتابعات باستخدام نوافذ متداخلة (منزلقة) ذات أطوال ثابتة، وكل مقارنة داخل المصفوفة لا بد أن تحقق قيمة حدية دنيا معينة مجموعة للنافذة ككل (الشرط) حتى يتسنى اعتبارها توافقا. فمثلا فى الشكل (٦-٩ب) نجد أننا قد استخدمنا حجم نافذة طوله ثلاثة أزواج من القواعد، ولا توضع نقطة فى المصفوفة إلا بشرط أن تتوافق قاعدتان على الأقل من الثلاث بين التتابعين، وبهذا تصبح المصفوفة أقل تشوشا.

(أ)

تتابع ١ (columns) — تتابع ٢ (rows)

	T	C	A	A	C	T	G	A	G	T	T	C	T	G	T	T	T	A	A	T	G
T	●					●				●	●		●		●	●	●			●	
C		●			●							●									
G							●		●					●							●
A			●	●				●										●	●		
C		●			●							●									
A			●	●				●										●	●		
G							●		●					●							●
G							●		●					●							●
G							●		●					●							●
T	●					●				●	●		●		●	●	●			●	
T	●					●				●	●		●		●	●	●			●	
T	●					●				●	●		●		●	●	●			●	
T	●					●				●	●		●		●	●	●			●	
G							●		●					●							●
C		●			●							●									
T	●					●				●	●		●		●	●	●			●	
T	●					●				●	●		●		●	●	●			●	
A			●	●				●										●	●		
A			●	●				●										●	●		
T	●					●				●	●		●		●	●	●			●	
G							●		●					●							●

(ب)

تتابع ١ (columns) — تتابع ٢ (rows)

	T	C	A	A	C	T	G	A	G	T	T	C	T	G	T	T	T	A	A	T	G
T																					
C		●										●									
G							●														
A				●				●													
C		●			●																
A			●			●		●					●								
G									●											●	
G								●													
G									●	●				●							
T										●					●	●					
T										●	●	●			●	●	●				
T										●	●	●			●	●	●				
T								●		●	●		●		●					●	
G							●							●							
C					●					●					●						
T						●				●			●		●	●					
T		●					●				●					●	●				
A			●														●	●			
A				●	●														●		
T						●							●							●	
G																					

(ج) تتابع ١

	S	T	E	F	C	L	M
S	•						
T		•					
G							
F				•			
C					•		
L						•	
M							•

شكل (٦-٩) — (أ) مصفوفة نقاط لتتابعين نيوكليوتيديين، الرصيص مغطى بالعديد من النقاط الزائفة فى المصفوفة. (ب) مصفوفة النقاط لتتابعى النيوكليوتيدات الموجودين فى (أ)، ولكن بعد استخدام حجم نافذة من ثلاث نيوكليوتيدات ووضع حد شرطية باثنين من ثلاثة توافقات لوضع نقطة فى العنصر المناسب، وتساعد عملية الغربلة هذه على حذف عدد كبير من النقاط الزائفة ومن ثم يظهر الرصيص بصورة أوضح على خلفية أقل تشوشا. (ج) مصفوفة نقاط لتتابعين من الأحماض الأمينية تم الحصول عليهما بترجمة تتابعى النيوكليوتيدات الموجودين فى (أ)، ويتضح كيف أن الرصيص لم يعد غامضا بالمرة.

وفى حالة التتابعات المشفرة للبروتينات، فإنه يفضل مقارنة تتابعات الأحماض الأمينية بدلا من مقارنة تتابعات الدنا، حيث أن زيادة الصفات (الحروف) من ٤ إلى ٢٠ ونقص طول التتابع من L إلى L/3 سوف يخفض بشدة من عدد النقاط الزائفة (شكل ٩ج).

٦. ٤. ٣. حسابات رص التتابعات.

لقد رأينا كيف أن استخدام طريقة مصفوفة النقاط لرص تتابعات ذات إحلالات وفجوات فإنها تعطى أكثر من رصيص، ولكن أيا منها هو الرصيص الأمثل optimal alignment، أى أفضل رصيص محتمل بين تتابعين، الذى يمكن الاعتماد عليه. والرصيص الأمثل هو ذلك المحتوى على الحد الأدنى من عدم التوافقات والفجوات تبعا لمعايير معينة. بيد أنه وللأسف، يؤدى خفض عدد عدم التوافقات إلى زيادة عدد الفجوات والعكس بالعكس.

فمثلا إذا كان لدينا التتابعان التاليان A و B :

$L_A = 11$ TCAGACGATTG A:

$L_B = 9$ TCGGAGCTG B:

فإنه يمكننا خفض عدد عدم التوافقات إلى صفر كما يلى :

TCAG-ACG-ATTG
(I)
TC-GGA-GC-T-G

ونحصل هنا على ٦ فجوات. وعلى النقيض من ذلك يمكننا خفض عدد الفجوات إلى فجوة واحدة لها أقصر طول مسموح به وهو هنا نيوكليوتيدتان حيث: $|L_A - L_B| = 2$، مما يترتب عليه زيادة فى عدد عدم التوافقات :

TCAGACGATTG
*   * * * *   (II)
TCGGAGCTG--

وفى هذه الحالة نحصل على فجوة واحدة (طولها نيوكليوتيدتان)، بيد أن عدد عدم التوافقات (المشار إليها بنجمة) يرتفع إلى ٥ (١ من نوع الـ transition و٤ من نوع الـ transversion).

وكبديل، يمكننا اختيار رصيص لا يخفض لا من عدد الفجوات ولا من عدد عدم التوافقات، مثل :

TCAG-ACGATTG
*   *   (III)
TC-GGA-GCTG-

وفى هذه الحالة يكون عدد عدم التوافقات ٢ (كلاهما من نوع الـ transversion) وعدد الفجوات ٤.

فمن إذن من بين الرصائص الثلاثة هو الأمثل ؟ ولأن مقارنة الفجوات مع عدم التوافقات تشبه مقارنة التفاح بالبرتقال أى لا معنى لها، ومن ثم فإنه يتحتم علينا إيجاد قاسم مشترك يمكن بواسطته مقارنة الفجوات بعدم التوافقات، وهو ما يطلق عليه غرامة الفجوة أو تكلفة الفجوة **gap penalty or gap cost**. وغرامة الفجوة هى عامل (أو عدة عوامل) تضرب بها قيم الفجوة (أعداد الفجوات وأطوالها) حتى تتساوى قيمة الفجوات مع

قيمة عدم التوافقات. وتبنى غرامة الفجوات على تقديرنا لتكرارات الأنواع المختلفة من الإدراج والحذف التى وقعت خلال التطور مقارنة بتكرار الإحلالات النيوكليوتيدية. وعليه فإنه لا بد لنا كذلك من تقدير غرامة عدم التوافقات mismatch penalties الذى هو تقدير لتكرار الإحلالات.

ولأى رصيص، يمكننا حساب مسافة أو مؤشر عدم تشابه distance or dissimilarity index بين التتابعين فى الرص ويرمز له بالرمز D

$$D = \sum m_i y_i + \sum w_k z_k \quad (٢٦)$$

حيث y_i عدد عدم التوافقات من النوع i، وm_i غرامة عدم التوافق من النوع i، وz_k عدد الفجوات ذات الطول k، وw_k رقم موجب يمثل غرامة الفجوة ذات الطول k.

وبالمثل، فإن التشابه بين تتابعين فى رص يمكن قياسه بمؤشر تشابه similarity index يرمز له بالرمز S، وهو يساوى

$$S = x - \sum w_k z_k \quad (٢٧)$$

حيث x عدد التوافقات.

وعادة ما يتم افتراض احتواء غرامة الفجوات على مكونين هما غرامة فتح الفجوة gap-opening penalty وغرامة إطالة الفجوة gap-extension penalty، وتعتمد الأخيرة على معلومية مسبقة بتكرار أحداث الحذف والإدراج ذات أطوال محددة بالنسبة لأحداث الإحلالات النيوكليوتيدية الأخرى. ففى نظام الغرامة الثابتة للفجوات fixed gap penalty system لا يتم حساب أى غرامة لإطالة الفجوات، أما فى نظام الغرامة المتصلة أو الخطية للفجوات affine or linear gap penalty system تحسب تكلفة إطالة الفجوة بضرب طول الفجوة مطروح منها ١ فى ثابت يمثل غرامة إطالة الفجوة بـ١. فمثلاً، بالنسبة لفجوة طولها ١ تتضمن تكلفة الفجوة غرامة فتح الفجوة فقط، أما بالنسبة لفجوة طولها ٣ فإن تكلفة الفجوة تتضمن غرامة فتح الفجوة بالإضافة إلى غرامة إطالة الفجوة مضروبة فى ٢. وحتى لا تزيد غرامة إطالة الفجوة بصورة كبيرة فى الفجوات الطويلة، فقد اقترح بعض العلماء مثل Gu and Li (1995) نظام الغرامة

اللوغاريتمية للفجوات logarithmic gap penalty system، وفيه تزداد غرامة الفجوة بصورة أبطأ مع الزيادة فى طول الفجوة.

وبالنسبة للرصائص الثلاثة السابقة، فإننا إذا استخدمنا غرامة عدم توافق تساوى ١ وغرامة فتح فجوة تساوى ٢ وغرامة إطالة فجوة تساوى ٦، يصبح مؤشر عدم التشابه للرصيص I يساوى $D = (0 \times 1) + (6 \times 2) + 6(1 - 1) = 12$، وبالمثل، فإن قيمة D للرصيصين II وIII تساوى 12 و10 على التوالى. وعليه فإنه من بين الرصائص الثلاثة هذه يعتبر الرصيص III هو الرصيص الأمثل (أقل D).

إن الغرض من أى حساب للرص alignment algorithm هو اختيار الرصيص ذى الـD الأقل (أو الـS الأعلى) من بين جميع الرصائص المحتملة الأخرى، والتى عادة ما تبلغ أعدادها أرقاما فلكية. فمثلا إذا تمت مقارنة تتابعين طول كل منهما ٣٠٠ فضالة فإننا نحصل على 10^{300} رص محتمل وذلك بالسماح بأى عدد وأى طول من الفجوات. ومن حسن الحظ فثمة حسابات كمبيوترية للعثور على الرصيص الأمثل لعل أشهرها حساب نيدلمان-وونش (1970)Needleman-Wunsch algorithm والذى يستخدم تقنية كمبيوترية عامة تدعى البرمجة الديناميكية dynamic programming وهى تقنية تستخدم حين يمكن تقسيم بحث كبير إلى عدد من الخطوات من المراحل الصغيرة بحيث يكون (١) حل مرحلة البحث المبدئية عاديا، (٢) كل حل جزئى للمراحل المتأخرة من البحث يمكن إيجاده بالرجوع إلى عدد قليل فقط من الحلول فى المراحل الأبكر، (٣) تحتوى المرحلة الأخيرة على الحل الكلى. ومن ثم يمكن إجراء البرمجة الديناميكية على قضايا الرص لأن مؤشرات التشابه تخضع للقاعدة التالية :

$$S_{1\to x,1\to y} = \max S_{1\to x-1,1\to y-1} + S_{x,y} \quad (٢٨)$$

حيث $S_{1\to x,1\to y}$ هى مؤشر التشابه بين التتابعين حتى الفضالة x فى التتابع الأول والفضالة y فى التتابع الثانى، و$\max S_{1\to x-1,1\to y-1}$ هو مؤشر التشابه لأفضل تتابع حتى الفضالة x – 1 فى التتابع الأول والفضالة y – 1 فى التتابع الثانى، و$S_{x,y}$ هو نتيجة التشابه لرص الفضالتين x وy.

وهو ما يصدق بحذافيره كذلك على مؤشر عدم التشابه بحيث :

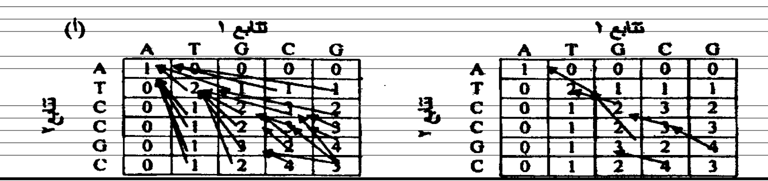
$$D_{1\to x,1\to y} = \max D_{1\to x-1,1\to y-1} + D_{x,y} \quad (٣٩)$$

ويتم حساب الرص على مرحلتين :

أولا، يتم رص التتابعين بطريقة مصفوفة النقاط، ولكل عنصر فى المصفوفة، وليكن x وy، يحسب مؤشر التشابه $S_{1\to x,1\to y}$. وفى نفس الوقت، يتم تخزين أفضل نتيجة للرص فى الصف أو العمود السابق، وتسمى القيمة المخزنة بالمؤشرة pointer. وتمثل العلاقة بين بين قيمة $S_{1\to x,1\to y}$ الجديدة والمؤشرة بسهم.

ثانيا، يتم إنتاج الرصيص بدءا بأعلى نتيجة تشابه سواء فى العمود الواقع أقصى اليمين أو فى الصف السفلى ثم يتم تعقب أفضل مؤشرة من اليمين إلى اليسار، ويطلق على هذه العملية القيافة traceback، كذلك يسمى رسم المؤشرات على خط القيافة برسم الطريق path graph لأنه يعرف الطرق عبر المصفوفة التى توافق الرصيص الأمثل.

ويظهر الشكل (٦-١٠) مثالا بسيطا لبرمجة ديناميكية لرص التتابعين ATGCG وATCCGC، ولتبسيط الأمر فإننا نفرض أن $w_k = 0$ (أى لا غرامة للفجوات)، ومن ثم فإن الرصيص الأمثل هو ذلك الرصيص ذو العدد الأكبر من التوافقات. وتملأ المصفوفة من اليسار إلى اليمين ومن أعلى إلى أسفل، ومن ثم نجد توافقا واحدا فى الصف الأول يحصل على النتيجة ١ وأربعة عدم توافقات كل منها يحصل على النتيجة صفر، أما فى الصف الثانى فإننا نبدأ بمزاوجة T من التتابع الأول بـT من التتابع الثانى وهو توافق ومن ثم فإننا نضيف ١ إلى أعلى نتيجة سابقة (وهى ١ فى الصف الأول) ونكتب المجموع عندها مع رسم سهم من هذا العنصر إلى مؤشرته، وهكذا دواليك إلى أن يتم الانتهاء من المصفوفة كلها (شكل ٦-١٠أ)، مع ملاحظة أنه يمكن لعنصر واحد أن يكون له أكثر من مؤشرة أو أن تشير أسهم أكثر من عنصر إلى مؤشرة واحدة. وفى المرحلة الثانية، نبدأ من أعلى نتيجتين فى آخر صف وآخر عمود من المصفوفة (مظللتين فى الشكل ٦-١٠ب) ونبدأ فى اقتفاء الرصيص بالبحث عن المؤشرة (أو المؤشرات) السابقة ذات القيمة الأعلى، حتى نحصل على طريق أو طرق تبرز الرصائص المثلى.



شكل (٦-١٠) — حساب البرمجة الديناميكية لرص على مرحلتين (أ وب)، ويلاحظ الحصول على رصيصين أمثلين فى (ب).

٦. ٥. رص التتابعات المتعددة.

٦. ٥. ١. حسابيات وبرامج الرصائص المتعددة.

يمكن اعتبار رص التتابعات المتعددة multiple sequence alignment امتدادا لرص التتابعات فى أزواج، بيد أن حساباته تزداد تعقيدا وبصورة أسية كلما زاد عدد التتابعات قيد الدراسة، ومن ثم يصبح من غير العملى إجراء بحث شامل عن الرصيص الأمثل فى حالة رص التتابعات المتعددة. ولقد تم نشر عدد من الطرق المساعدة على ذلك، كما تم وضع عدد من برامج الكمبيوتر المتاحة على الانترنت التى تقوم بالعثور على الرصيص الأمثل لتتابعات متعددة (أنظر القائمة التالية):

1- ***Baylor's search launcher for Biologists***, Multiple Alignment section
ClustalW, MAP, PIMA 1.4, MSA 2.1, and BLOCK MAKER alignments
http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html

2- **Clustalw + Multalin** Multiple Alignment at IBCP
Calculate, optimize and score multiple sequence alignments with **ProbModel, MSA, ClustalW** and heuristics. The scoring is a probabilistic measure.
http://cbrg.inf.ethz.ch/MultAlign.html

3- **STRAP**: Interactive program for generating and analyzing multiple sequence alignments. Multiple structure alignments, integrated 3D-viewer, 3D-superposition of protein backbones, mapping of mutations

onto 3D-models, translation of nucleotide sequences to amino acid sequences, sequence dotplots.
http://www.charite.de/bioinf/strap/

4- **MUSCA**: An Algorithm for Constrained Alignment of Multiple Data Sequences, and related tools.
http://cbcsrv.watson.ibm.com/Tmsa.html

5- **JEvTrace** - algorithms and multivalent graphical browser for combined alignment, phylogeny, and structure analysis.
http://www.cmpharm.ucsf.edu/~marcinj/JEvTrace/

6- **MAVID** - multiple alignment program for large genomic regions
http://baboon.math.berkeley.edu/mavid/

7- **MUSEQAL** optimal multiple alignment by iteratively improving a given set of pre-aligned sequences.
http://godzilla.dcrt.nih.gov/~yap/museqal2.html

8- **DCA**: Divide and Conquer Multiple Sequence Alignment
Close-to-optimal, fast simultaneous sum-of-pairs alignment of up to twenty protein, DNA, or RNA sequences
http://bibiserv.techfak.uni-bielefeld.de/dca/

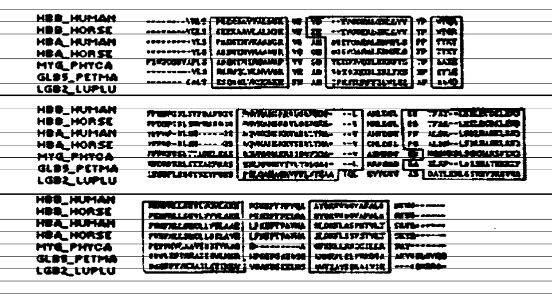
9- **DIALIGN** - fragment-based alignment program following the method described by Morgenstern *et al.* in PNAS 93 (22), pp. 12098-12103.
http://cartan.gmd.de/ToPLign.html

10- **ToPLign**: Toolbox for Protein Alignment
Computing, analysis and visualization of pairwise, multiple, threading, and parametric alignments.
http://www.ibc.wustl.edu/msa.html

11-**MSA**, (Close-to-) Optimal Alignments using the Carrillo-Lipman bound
http://www.toulouse.inra.fr/multalin.html

لعل أشهرتلك البرامج وأكثرها استخداما هو برنامج CLUSTAL الذى وضعه Higgins and Sharp (1988, 1989) والذى سوف نتناوله بالشرح فيما بعد. وتستخدم معظم هذه البرامج حسابا تدرجيا أو اطراديا incremental or progressive algorithm (Feng and Doolittle, 1987)، وفيه يضاف تتابع جديد لعدد من التتابعات المرصوصة أصلا بترتيب الأقل تشابها. وتبدأ العملية بإجراء كافة الرصوص الممكنة بين أزواج التتابعات قيد الدراسة وإيجاد الزوجين الأكثر تشابها بينها، وبناء عليه يبدأ اختيار التتابعات التالية وهلم جرا.

ويوضح الشكل (٦-١١) مثالا لرصيص من التتابعات المتعددة، وهى نخبة صغيرة من تتابعات الجلوبين ذات تراكيب ثلاثية معروفة، حيث أن لكل منها نفس التثنى folding الذى اكتشفه Kendrew and Perutz فى الخمسينيات الذى يتكون من فضالتين resudues محفوظتين من الهستيدين تربط مجموعة الهيم الترقيعية prosthetic بقلب core محفوظ من سبعة لوالب "ألفا" تتشارك فيها كل هذه البروتينات. وعلى طول التاريخ التطورى، أخذت الفضالات تستبدل بأخرى (عادة مع الاحتفاظ بالخصائص الكيموحيوية لسلاسل الأحماض الأمينية الجانبية) مما دفع التتابعات للتشعب عن بعضها البعض، كما هو مقاس بالنسبة المئوية للتشابه بينها. وفى حالتنا هذه، فإن أقل تتابعين



شكل (٦-١١)، رصيص متعدد لسبع تتابعات من بروتينات الجلوبين؛ (١) جلوبين بيتا الإنسان (٢) جلوبين بيتا الحصان (٣) جلوبين ألفا الإنسان (٤) جلوبين ألفا الحصان (٥) ميوجلوبين الحوت (٦) سيلنوهيموجلوبين المورَيل البحري و(٧) لجهيموجلوبين نبات الترمس.

تشابها هما تتابعى ميوجلوبين الحوت (الثالث من أسفل – رقم ٥) ولجهيموجلوبين نبات الترمس البقولي (التتابع الأخير- رقم ٧)، حيث لا يتعدى التماثل بينهما أكثر من ١٠٪ ومع ذلك فإن التشابه بينهما يظل ظاهرا عند رصهما.

وتجدر الإشارة إلى أنه للحصول على مثل هذا الرصيص فلا بد أولا من التأكد من صحة ودقة الحصول على البيانات الجزيئية (التتابعات) حيث أنه ما من برنامج يقدر على إنتاج رصيص صحيح من بيانات غير صحيحة، وهى مهمة يضطلع بها باحثو البيولوجيا الجزيئية قبل استخدام برنامج الكمبيوتر، حيث عليهم توخى الحذر والنظر إلى التتابع بغرض حذف الأجزاء المسئولة عن تشويه النتائج. ويمكن تحقيق هذا باتباع مدخل iterative، فيقوم الباحث بعمل عدة رصائص اختبارية ويختبر كلا من الشجرة الفيلوجينية المتحصل عليها علاوة على الرصائص البينية.

وأخيرا، إذا توفرت للباحث معلومات إضافية مثل معرفة فضالات موقع نشط محدد فإن مثل هذه المعلومات من شأنها أن تساعد على إلقاء الضوء على بعض أخطاء الرص أو على التحرير اليدوى للرصيص. وهو ما يجعل الباحث فى حاجة إلى استخدام نوعين من البرامج أحدهما ليولد رصيصا مناسبا والآخر لتحريره.

وعادة ما يتم اختيار التتابعات من قواعد البيانات الموجودة على الانترنت مثل SwissProt (Bairoch and Apweiler, 1996). باستخدام برامج البحث على أساس التشابه مثل BLAST (Altschul *et al.*, أو FASTA (Pearson and Lipman, 1988) (1997. وكذلك يمكن إجراء البحث باستخدام مفاتيح keywords أو أدوات البحث داخل قواعد البيانات مثل SRS (Etzold *et al.*, 1996) أو كليهما. والطريقة الأولى دوما ما تنتج تتابعات شبيهة بالتتابع قيد الدراسة (والذى قد تم الحصول عليه معمليا أو من احد الأدبيات) حتى وإن كان هذا التشابه لا يعكس أى علاقة تطورية وإنما هو وليد تركيب الأحماض الأمينية الموجودة فيهما، ومن ثم فعلى الباحث إعمال عقله ومعرفته بالخلفية البيولوجية للتابع قيد الدراسة حتى يتسنى له اختيار التتابع الصحيح للرص.

وفى الحالة الأولى، قد يساعد رص التتابعات المتعددة الباحث فى معرفة ما إذا كان أى توافق هو توافق صحيح وليس وليدا للصدفة، فكما يظهر فى الشكل (٦-١١) فإننا نجد أنه ثمة عدة فضالات محفوظة بين كل الجلوبينات فى الرصيص (مثل هستيدينى ربط الهيم)، حيث تبدو هذه الفضالات كحزم رأسية محفوظة وهو ما لا يمكن اكتشافه عند رص تتابعين فقط، فنحن مثلا لا نستطيع البت بوجود تشابه من عدمه بين ميوجلوبين الحوت ولجهيموجلوبين الترمس عند رصهما وحدهما، بيد أن هذا التشابه يظهر جليا عند استخدام رص التتابعات المتعددة. ولكن لا بد من الانتباه إلى إذا ما كانت التتابعات المتحصل عليها باستخدام التشابه تحتوى على شظايا قصيرة كمجالات غير تامة أو على مجالات متعددة كالفيبرونكتين والـ SH3 حيث أن هذا سوف يسبب مشاكل بالنسبة لصحة الرصيص ومن ثم يفضل رص المجالات كل على حدة إلا إذا احتوت كل التتابعات على نفس المجالات بنفس الترتيب.

أما فى الحالة الثانية عند استخدام مفاتيح بحث بدلا من تشابه التتابعات، فلا بد من معرفة أن وجود تتابعين لهما نفس مفتاح البحث لا يعنى بالضرورة وجود علاقة تطورية بينهما. فعلى سبيل المثال قد تقوم بروتينات مختلفة بنفس الوظيفة الإنزيمية فى مجموعات الكائنات الحية المختلفة، أو يكون مصحح قاعدة البيانات أو العالم نفسه قد قاما بخطأ عند تسمية التتابع.

٦.٥.٢. استخدام برنامج CLUSTAL W فى رص التتابعات.

تكمن قيمة برنامج CLUSTAL W فى رص التتابعات المتعددة للدنا والبروتينات فى أنه يعطى رصائص ذات دلالات بيولوجية للتتابعات العديدة المتشعبة من بعضها البعض، حيث يقوم بحساب أفضل توافق للتتابعات المختارة ثم رص هذه التتابعات بصورة يمكن معها مشاهدة التشابهات والاختلافات بينها وذلك بتعيين المناطق المحفوظة فيها، وهو ما يعد خطوة أساسية عند تصميم التجارب لاختبار وتحوير وظيفة بروتين ما أو عند التنبؤ بتركيب البروتينات ووظيفتها وكذلك عند تعيين أعضاء جدد فى العائلات البروتينية. كذلك يقوم البرنامج بتوضيح العلاقات التطورية بين التتابعات فى صورة شجرة فيلوجينية phylogram أو طائفية cladogram.

وبرنامج CLUSTAL W متاح على عدة مواقع على الانترنت منها http://www.ebi.ac.uk/clustalw/. وثمة طريقتان لاستخدامه على هذا الموقع، الأولى تفاعلية interactive، حيث ينتظر المستخدم ظهور النتائج على نافذة الاستعراض browser window، أما الطريقة الثانية فهى عن طريق البريد الالكترونى email حيث يتم إرسال النتائج على البريد الالكترونى للمستخدم، وتعتبر الطريقة الثانية هى الفضلى عند رص عدد كبير من التتابعات. كذلك يمكن الحصول على CLUSTAL W بطرق اخرى.

ويمكن إدخال التتابعات إلى البرنامج بأحد سبعة تنسيقات وإن كانت Fasta وALN/ClustalW هما أكثرها استخداما، أما نتائج الرص فيحصل عليها بتنسيق ALN/ClustalW أى بالامتداد.aln، وهو النسق الذى يتيح تلوين الرصيص فى حالة البروتينات تبعا لخواصها الكيموفيزيائية على النحو الموضح فى الجدول التالى:

AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RHK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Gray	

ويستخدم هذا الجدول عند كتابة الرصيص، وكذلك يدل الرمز "*" على التوافق، والرمز ":" على حدوث إحلال محفوظ، والرمز "." على لإحلال نصف المحفوظ.

٧. الجينومكس المقارن و الفيلوجينيا

Comparative Genomics & Phylogenetics.

إعداد: سناء رياض و أحمد المتينى

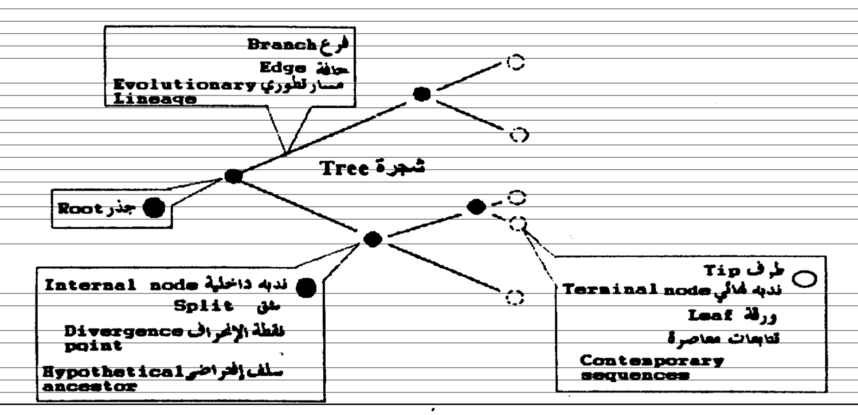
إن الدراسات المبنية على تتبع علاقات القرابة بين الكائنات (phylogenetic) اعتمادا على تحليل الـ DNA أو البروتينات، اكتسبت أهمية فى السنوات القليلة الماضية بغرض دراسة العلاقات التطورية للكائنات من أول البكتيريا حتى الإنسان. وحيث أن معدل التغير لهذه التتابعات (على مستوى الـ DNA أو البروتين) عالى وكبير بشكل ملحوظ (Nei, 1996)، فإننا نستطيع فى حقيقة الحال دراسة العلاقات التطورية على كافة المستويات التقسيمية للكائنات (kingdom, phyla, classes, families, genera, species, and populations). ولتعدد هذه الوحدات التقسيمية، لجأ العلماء لتحديدها (على إختلاف أشكالها) بمصطلح الوحدة التقسيمية (taxon)، والتى قد تكون كائنا أو جينا أو حتى تتابع من الـ DNA أو من البروتين، والتتابعات الجزيئية الأخيرة ستمثل أساس تناولنا لهذا الموضوع حيث أنها أكثر إرتباطا بمفهوم المعلوماتية الحيوية. كذلك فإن مثل تلك الدراسات يمكن أن تلقى الضوء على تطور العائلات الجينية (gene families) ويمكنها أيضا أن تلقى ببعض الضوء على ميكانيكيات الحفاظ على تعدد الأشكال المظهرية (polymorphic genes) على المستوى الجزيئى.

٧.١. الشجرة التطورية Evolutionary tree.

إن الشجرة التطورية أو شجرة القرابة (phylogenetic tree) تعرف أيضا بأسم dendrogram ويعنى العلاقة بين مجموعة من الأفراد المختلفة وإذا كانت تلك العلاقات فيولوجينية تسمى العلاقة cladogram. وقد بدأت تلك الدراسات منذ ٤٠ سنة مع بدايات علم التقسيم الرقمى (numerical taxonomy) المعتمد على الصفات المظهرية التقليدية

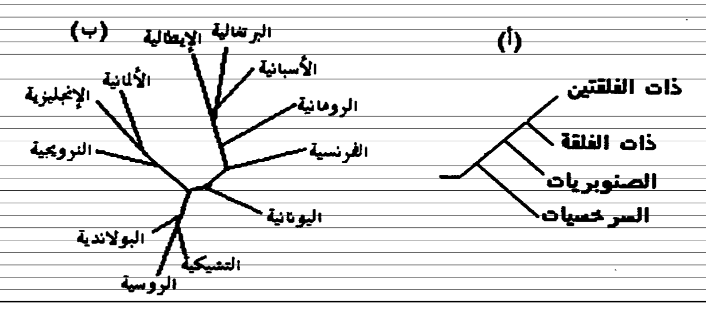
وعلى دراسة التكرارات الجينية فى العشائر، وبعض من برامج التحليل التى وضعت فى السابق لهذه الأغراض مازالت تستعمل حتى اليوم لتحليل البيانات الجزيئية، وأن كان لدينا الآن برامج أكثر حداثة وكفاءة لهذا الغرض.

والتقنية العملية لتصميم شجرة القرابة، تتبع الأساليب الإحصائية لعمل شجرة اعتبارية ممثلة للشجرة الحقيقية، ولهذا فلابد من معرفة طوبوغرافية (topology) الشجرة او نظام تشعبها أولا ثم يتبع ذلك حساب مدى أو طول هذا التشعب. والشجرة عبارة عن رسم ثنائى الأبعاد من فروع (branches) تمثل درجات القرابة بين الـ taxa، والتى قد تكون مرتبطة بمنشأ مشترك واحد يسمى عادة بالجذر root تخرج منه الفروع branches الممثلة للمسارات التطورية ويوجد عليها نذب إما داخلية internal nodes وإما طرفية نهائية terminal nodes وهى تمثل البراعم الجانبية والبراعم الطرفية للشجرة والتى يمكن إعتبارها الأوراق leaves. وشكل (٧-١) يمثل شجرة قياسية مبين عليها المصطلحات والمرادفات المستخدمة.



شكل (٧-١) : رسم توضيحى لشجرة فيلوجينية قياسية.

وقد تكون الشجرة ذات جذر (منشأ) rooted tree كما هو مبين بشكل (٧-٢أ) أو عديمة الجذور un-rooted tree وهو ما يعرف أيضا بأسم الشجرة النجمية star tree كما هو مبين بشكل (٧-٢ب).



شكل (٧-٢) : (أ) مثال للشجرة ذات الجذور للعلاقة بين النباتات الراقية، (ب) مثال للشجرة النجمية للعلاقة بين اللغات الأوربية.

٧.٢. طرق تعيين الشجرات Tree determination methods.

إن بناء طوبوغرافية الشجرة، يتطلب قدرا كافيا من البيانات الجزيئية (افترض انها من عشر مصادر) لذا فإن عدد تفرعات الشجرة المحتملة قد يصل إلى المليون وعليك أن تختار منها الأنسب. ولقد اقترحت لذاك عشرات الطرق الإحصائية المعتمدة على الكمبيوتر لتحقيق ذلك ويمكن إجمالها تحت ثلاث أقسام رئيسية هى :

١- - الطرق المقتصدة parsimony methods.
٢- طرق حساب المسافة (البعد) distance methods
٣- طرق حساب ارجحية التشابه likelihood methods.

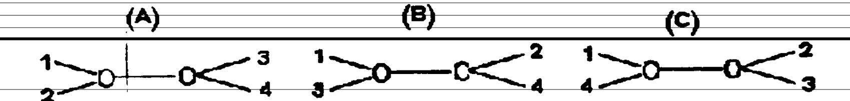
٧.٢.١. الطرق المقتصدة Parsimony Methods .

إن الطرق المقتصدة لعمل الشجرات التطورية بين مجموعة من التتابعات تعتمد أساسا على إمكانية اختيار العدد الأدنى من التغيرات فى التتابعات بين الـ taxa تحت الدراسة. ولعمل الشجرة بأتباع هذه الطريقة لابد أولا من عمل الصف المتعدد

(multiple alignment) للعينات تحت الدراسة، والشجرة التى يمكن أن تفسر الإختلافات الموجودة عند أقل عدد ممكن من التبادلات هى التى يتم اختيارها. وهذه الطرق مطولة وتحتاج لمجهود ووقت كبير، وعادة تنجح فى دراسة التتابعات التى بينها درجات عالية من القرابة. ولتوضيح هذه الطرق دعنا نناقش المثال التالى، فالتتابعات التالية تمثل أربعة عينات مختلفة وتتابعاتها هى:

1 AAGAGTGCA
2 AGCCGTGCG
3 AGATATCCA
4 AGAGATCCG

هذه التتابعات الأربع يمكن أن تكون فيما بينها ثلاث شجرات نجمية (ليس لها جذر) كما هو موضح بشكل (٧-٢).



شكل (٧-٢) : الشجرات النجمية الثلاث التى يمكن ان تمثل العلاقات بين التتابعات الأربع المبينة.

و يجب أن نلفت النظر إلى ملاحظة هامة وهى أن التتابعات السابقة يوجد بها صفوف مفيدة معلوماتيا informative وأخرى غير معلوماتية un-informative. الصف المعلوماتى هو الموضع الذى يلاحظ به أثنين من القواعد المتماثلة على الأقل، وعلى ذلك فجميع المواضع السابقة معلوماتية ما عدا الموضع الرابع حيث لا توجد قواعد متماثلة على الإطلاق. وعلى البرنامج الحسابى (برنامج الكمبيوتر) أن يميز الشجرة الأكثر كفاءة منها حيث يوجد أقل عدد ممكن من التغيرات بين التتابعات الأربع. وعلى ذلك يمكن القول بأن الشجرة الأولى (A) هى الأكثر كفاءة لتمثيل العلاقة بينهم حيث عدد التغيرات سيكون ٣ و ١ فقط عند مواقع التلاقي على الشجرة، بينما النماذج الأخرى ستعطى عدد أكبر من التغيرات.

وهذه الطرق تعتبر من أقدم الطرق المستخدمة لتقدير شجرات القرابة لكن كفاءتها محدودة خصوصا عند التعامل مع عدد كبير من الوحدات التقسيمية مستخدمين تتابعات كبيرة للصف والمقارنة.

لكن يجب التنويه بأن البرامج الحسابية أو برامج الكمبيوتر المختصة بتقدير وتصميم الشجرات التقسيمية التطورية تعتمد فى بعض منها على نظرية من نظريات علم وراثة العشائر population genetics الا وهى نظرية "الزمن الجزيئى" molecular clock التى ظهرت فى أواخر الستينات من القرن الماضى. فمن دراسات تغيرات (طفرات الإستبدال) الأحماض الأمينية فى البروتينات، إتضح أن معدل الإستبدالات ثابت تقريبا مما يعنى أن التفرعات على الشجرة إذا ما تساوت فى أطوالها branch length يعنى التساوى فى عدد الإستبدالات المحتملة وعليه يمكن تحديد زمن التفرع على الشجرة أو زمن الفصل بين الوحدات التقسيمية. وقد لاقت هذه النظرية قبولا كبيرا بين علماء وراثة العشائر والتطور الذين يؤمنون بحيادية الإختلافات الجزيئية neutrality وعلى رأسهم Kimura، وإن لاقت عدم قبول من بعض العلماء المتشككين من مبدأ الحيادية هذا. وعليه فهناك برامج حسابية algorithm تأخذ فى الحسبان نظرية الزمن الجزيئى وغيرها تتجاهلها.

٢.٢.٧. طرق تقدير البعد Distance methods.

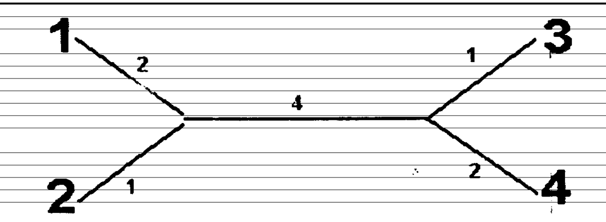
هناك العديد من الطرق التى تعتمد على تقدير البعد الوراثى بين الـ taxa تحت الدراسة، وسنحاول هنا تناول أهمها بالشرح والتوضيح. فلعمل الشجرات إعتمادا على البعد، تقدر العلاقات بين العينات إما بحساب عدد الإستبدالات على الفرع أو بحساب عدد التغيرات الواجب إجرائها لجعل أى تتابعين متناظرين مرة أخرى. ونجاح هذه الطرق يعتمد على قدرتها فى تبسيط الإختلافات وتحويلها إلى صورة قيم مضافة additive. ولتوضيح الفكرة الأساسية من تلك الطرق دعنا ندرس التتابعات التالية بين أربع من العينات لعمل شجرة تربطهم:

1. ACGCGTTGGGCGATGGCAAC
2. ACGCGTTGGGCGACGGTAAT
3. ACGCATTGAATGATGATAAT
4. ACACATTGAGTGATAATAAT

وبحساب البعد بين هذه التتابعات الأربع نحصل على النتائج التالية:

	1	2	3	4
1	-	3	7	8
2	-	-	6	7
3	-	-	-	3
4	-	-	-	-

وباستعمال هذه المعلومات يمكن تصميم شجرة قرابة نجمية (عديمة الجذر) بين التتابعات الأربع حيث أن قيم الإستبدالات (الإختلافات) بين أزواج العينات مقدرة كأطوال فروع الشجرة كما هو مبين كالتالي:

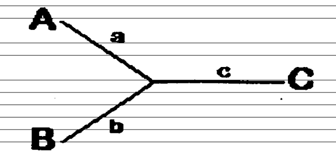


وعادة ما يفضل كثير من الباحثين التفكير فى الشجرات التطورية على أن لها سلف مشترك أو على أنها شجرة لها جذر rooted، لذلك لجأت كثير من الطرق لحسابات خاصة تحقق ذلك.

٧. ٢. ٢. ١. طريقة Fitch & Margoliash.

وهى من أوائل الطرق التى أقترحت فى هذا الصدد فى أوائل السبعينات من القرن الماضى والتى تحدد الوحدات داخل الشجرة فى ثلاثيات threes ثم تحسب قيم البعد لأذرع تلك الشجرة. وفيما يلى تعريف بأسس هذه الطريقة:

[١] تقترح شجرة من ثلاث وحدات كالتالى:



[٢] تحسب أطوال الأذرع جبريا بناء على درجات البعد والمبينة كالتالى:

	A	B	C
A	-	22	39
B	-	-	41
C	-	-	-

Distance from A to B = a + b = 22 .... (1)

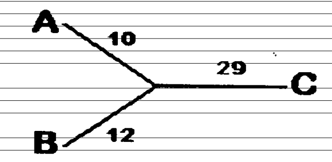
Distance from A to C = a + c = 39 .... (2)

Distance from B to C = b + c = 41 .... (3)

للتخلص من المجاهيل تطرح (2) من (3) فيكون الناتج

(b + c) – (a + c) = 41 – 39 = b – a = 2 .... (4)

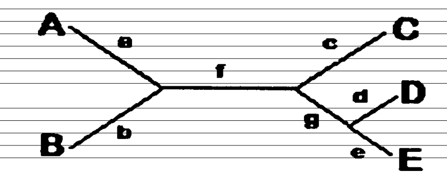
بالجمع (1) و (4) الناتج سيكون 2b = 24; b = 12 وبالتعويض فى (1) و (2) تكون قيم a = 10 و c = 29 . وعلى ذلك تمثل الشجرة كالآتى:



ولكن ما هو الحال إذا ما كان عدد العينات أكثر من ثلاث، مثلا خمسة A, B, C, D, E ، وعلاقات البعد بينهم يلخصها الجدول التالى:

	A	B	C	D	E
A	-	22	39	39	41
B	-	-	41	41	43
C	-	-	-	18	20
D	-	-	-	-	10
E	-	-	-	-	-

ويمكن تمثيل الشجرة كالتالى:



فيتم التعامل معها كالتالى:

[١] اولا يتم تحديد اكثر أثنين من العينات تحت الدراسة قربا من بعضهما البعض من جدول البعد الوراثى وفى هذه الحالة ستضح أنهما العينات D و E

[٢] نعيد حساب درجات البعد على انها ثلاث عينات، فلحسلب بعد D عن A, B, C مجتمعة يأخذ متوسط البعد لهم الثلاث عن D، أى (39 + 41 + 18)/3 = 32.7 وللبعد بين E و A, B, C وبإتباع نفس الطريقة سيكون 34.7 ، وعليه نعدل الجدول ليصبح

	D	E	Ave. A,B,C
D	-	10	32.7
E	-	-	34.7
Ave. A,B,C	-	-	-

[٣] ويمكن ايضا حساب العلاقة بين D و A,B,C وكذلك بين E و A,B,C عن طريق حساب متوسطات أطوال الأفرع على الشجرة:

D to E: d + e = 10 .... (1)

D to ABC: d + m = 32.7 .... (2)

where "m" = g + (c + 2f + a + b)/3

E to ABC: e + m = 34.7 .... (3)

وبالطرح المعادلة (٣) من المعادلة (٢) فإن d – e = -2 وبالجمع مع المعادلة (١)

نحصل على أن 2d = 8; d = 4 وبالتعويض فإن e = 6 .

[٤] والآن نتعامل مع العينات D و E على أنها عينة واحدة ونبنى جدول جديد لقيم البعد كالآتى:

	A	B	C	(DE)
A	-	22	39	40
B	-	-	41	42
C	-	-	-	19
(DE)	-	-	-	-

ومنها يمكن حساب قيم c ستساوى 9 وقيمة g ستساوى 5.

[٥] نعيد الخطوات السابقة حتى نحدد جميع قيم أطوال الأفرع على الشجرة تحت الدراسة.

<u>ملخص خطوات طريقة Fitch-Margoliash :</u>

١- حدد أكثر اثنين من العينات قربا.

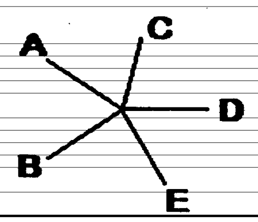
٢- اعتبر باقى العينات وحدة واحدة، ثم إحسب متوسط البعد بينهما ومتوسط العينات المجمعة.

٣- استعمل تلك العلاقات الجديدة فى حساب أطوال أفرع الشجرة لهما.

٤- اعد الخطوات السابقة مع تجميع عينات اخرى ومنها إحسب أطوال الأفرع وهكذا.

٧.٢.٢.٢.طريقة حساب ربط المتجاورات Neighbor-joining algorithm

ان هذه الطريقة قريبة الشبه جدا بالطريقة السابقة، فالعينات (التتابعات) التى يتم إختبارها للوصل بينها تعتمد على تقدير مربع الإنحرافات least-square لأفضل التقديرات لطول أفرع الشجرة. وتبدأ الطريقة بتحديد الشجرة النجمية حيث لا توجد اى إتصالات بين العينات المتجاورة، كما هو موضح بالشجرة التالية:

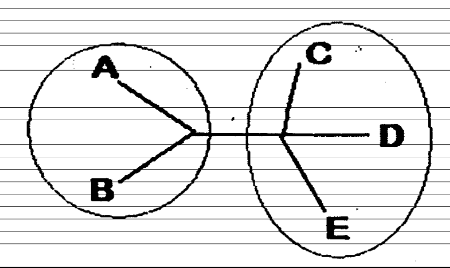


وتحديد اى من العينات سيتم ربطها معا عن طريق حساب مجموع أطوال الأفرع للشجرة بناء على المعادلة:

$$S_{mn} = \frac{\sum d_{im} + d_{in}}{2(N-2)} + \frac{d_{mn}}{2} + \frac{\sum d_{ij}}{N-2}$$

where ij represent all sequences except m and n, and i < j.

وعلى سبيل المثال دعنا نقول أن التتابعات (العينات) المتجاورة التى ستوصل أولا هى A و B، كما هو مبين بالشجرة التالية:



٧.٢.٢.٢.٣. طريقة UPGMA.

وهذه الطريقة هى أكثر الطرق إستعمالا لتقدير البعد و تحديد الشجرات التطورية مستخدمة تتابعات الـ DNA والبروتين. و هذا الأسم UPGMA هو إختصار لطريقة الحساب المعروفة بأسم Un-weighted Pair Group Method with Arithmetic Mean. ويبدأ الحساب بمحاولة تجميع clustering التتابعات فى أزواج (الأقرب لبعضهما) و ربطهما مع بعضهما فيما يسمى بالندبة او النتوء node ثم زوج ثانى ثم ثالث وهكذا وفى ذلك نقوم ببناء الشجرة من أسفل إلى أعلى upwards فالنتوء الأول يعلوه الثانى والثانى يعلوه الثالث وهكذا وحواف (حدود) النتوء يحسب من فرق ارتفاع هذه الوحدات (النتوءات).

والمسافة بين كل تجميعتين clusters مثلا C_i و C_j تحسب من المعادلة التالية:

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p in C_i, q in C_j} d_{pq}$$

where $|C_i|$ and $|C_j|$ are the number of sequences in clusters i and j, respectively

<u>وحسابات البرنامج لطريقة UPGMA algorithm يمكن تلخيصها فى الآتى:</u>

[١] اربط كل تتابع i لمجموعته C_i cluster.

[٢] حدد كل ورقة (وحدة) من أوراق الشجرة T لكل تتابع، و وقعها عند الارتفاع صفر.

[٣] حدد المجموعة من i و j بإستعمال المعادلة السابقة التى تعطى أقل تقدير لقيمة d_{ij} .

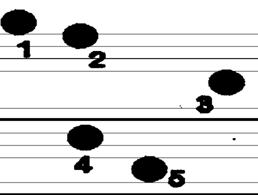
[٤] حدد مجموعة ثانية k حيث $C_k = C_i \cup C_j$ ثم حدد d_{kl} مع كل التتابعات.

[٥] حدد النتوء k بالنسبة للنتوء ين i و j ، و وقعها على إرتفاع يساوى $d_{ij} / 2$.

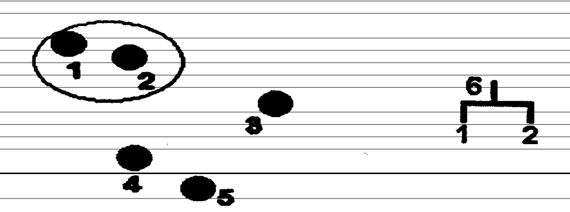
[٦] اضف k للمجموعة وهكذا يستمر العمل لباقى التتابعات.

<u>مثال تطبيقي على طريقة UPGMA :</u>

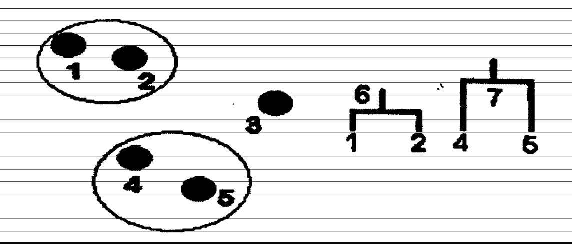
إفترض وجود ٥ تتابعات يراد تحديد العلاقات بينها، وممثلة كنقاط على رسم بيانى يحدد اماكنها، كالتالى :



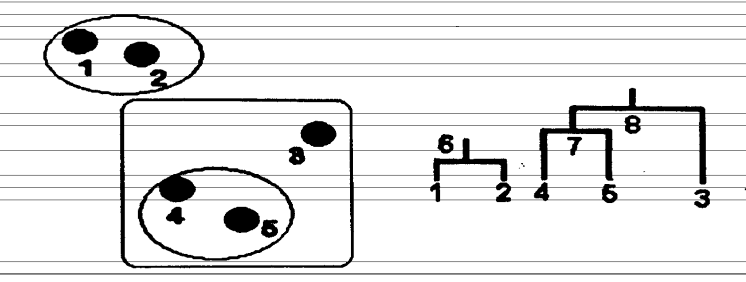
الآن إختار المجموعتين clusters الأكثر قربا من بعضهما البعض، وهما التتابعان ١ و ٢، كون مجموعة واحدة منهما (كما هو موضح بالشكل التالى) ثم انشأ نتوء node يمثلهما عند أرتفاع $d_{12} / 2$:



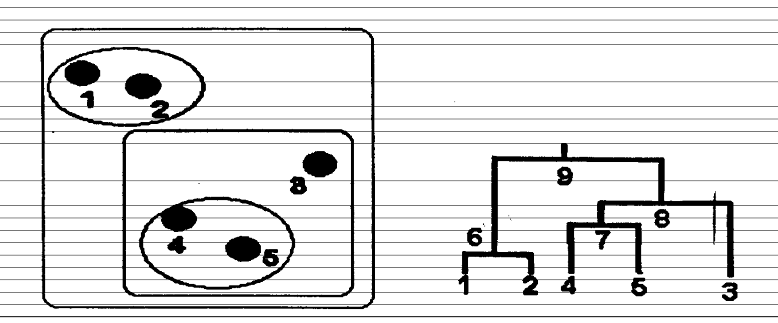
تابع العمل فإختار المجموعتين التاليتين من حيث القرب، وهما ٤ و ٥ وضمهما فى مجموعة واحدة، وتابع تكوين الشجرة، كما هو مبين كالتالى:



والمجموعة التالية هى التى سترتبط بين مجمعة (٥/٤) ومجموعة رقم ٣، كالتالى:

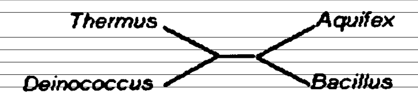


الآن لم يتبقى سوى مجموعتين لذا يجب الربط بينهما للإنتهاء من عمل الشجرة كالتالى:

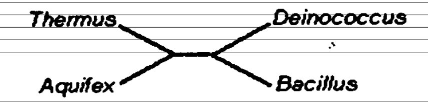


طرق تحديد الشجرات الفيلوجينية التى تناولناها بالشرح فى الأجزاء السابقة ومع أنها مستعملة على نطاق كبير فى التجارب الوراثية تعطى نتائج لابأس بها إلا أن كفاءتها محدودة خصوصا فى الدراسات التطورية فعلى سبيل المثال هناك أربعة أقسام من البكتيريا معروفة وراثيا وتطوريا بإختلافها وهى:

١- قسم (أى وحدة تقسيمية) *Deinococcus*، وهى بكتيريا مقاومة للأشعة radiation-resistant و -٢- قسم *Thermus* وهى محبة للحرارة thermophilic . وهناك أدلة تجريبية قوية بأن هذان القسمان متقاربان وراثيا و -٣- القسم *Aquifex* وهى محبة أيضا للحرارة و -٤- القسم *Bacillus* وهى محبة للماء mesophilic، والقسمان الثالث والرابع لاتربطهما أى علاقة قرابة بالقسمين الأول والثانى، وعلى ذلك كان المتوقع أن تكون الشجرة بينهم كالتالى:



وبإستعمال معظم الطرق السابقة ونتيجة التشابه فى تتابعات الـ DNA بين *Aquifex* و *Thermus* فأنهما يقعان متجاورتان على الشجرة خلافا للمتوقع كالآتى:



وللتغلب على تلك المشاكل يلجأ العلماء إلى طرق حسابية أخرى لتحديد الشجرة وتعرف طرق الأرجحية العظمى maximum likelihood methods، حيث أن هذه الطرق تعتبر أكثر كفاءة خصوصا عند التعامل مع الأفرع الطويلة للشجرات والبيانات الغريبة مثل التتابعات الشاذة وعند إنحراف نسبة الإستبدالات transitions/transversions عن الواحد أو عند إختلاف معدل التحول فى منطقة من التتابع عن مجاوراتها بمعنى أن يكون معدل التطور بطئ فى منطقة وسريع فى منطقة أخرى. مع العلم أن طرق الأرجحية قد تؤدى لعمل شجرة خاطئة كالمثال السابق إلا أنها تختلف عن الطرق السابقة فى قدرتها على عرض الشجرات الأقل أرجحية ولا ترفضها تماما.

٣.٢.٧. طرق الأرجحية العظمى Maximum Likelihood Methods.

طرق الأرجحية العظمى هى طرق تعتمد على تقدير الاحتمال نسبة لنموذج حسابي فرضى محدد model. فعلى سبيل المثال إذا ما ألقينا بقطعة نقود عادية (النموذج يحدد أن لها وجهان — الصورة والكتابة) فإن أرجحية الحصول على الصورة من رمية واحدة هى 0.5. و إذا نص النموذج أن قطعة النقود هذه غير طبيعية وعلى وجهيها صورتين، فإن الأرجحية فى هذه الحالة ستساوى الواحد الصحيح، أي أن طرق الأرجحية تعتمد على تقدير الاحتمال وفقا لنموذج مفترض. وفى تصميم النموذج model تبعا

لطرق الأرجحية لتعيين شجرة فيلوجينية لتتابعات عينات ما فإن النموذج يجب ان يشتمل على جزئين وهما:

١- المكون composition، ويختص بنسب النيوكلوتيدات الأربع (A,C,G,T) لبعضها البعض و ٢- النهج process ، ويختص بمعدلات الاستبدالات substitutions بين هذه النيوكلوتيدات الأربع. وفيما يلى تعريف مبسط بالأساس الحسابى لبرامج تقدير الشجرات بطرق الأرجحية العظمى.

فلحساب ارجحية تتابع ما (قل نيوكلوتيدة واحدة من A للتبسيط) فهنا لن توجد شجرة وسنستبعد أيضا جزء النهج process من النموذج وسيتوقف فقط على جزء المكون component، فإذا كان النموذج ينص على أن التتابع ١٠٠٪ A إذن الأرجحية ستكون واحد، وإذا ما نص النموذج بأنه ١٠٠٪ C إذن الأرجحية ستكون صفر وإذا ما نص النموذج بأن A تساوى ٣٣٪ فإن الأرجحية ستساوى ٠,٣٣.

وما هى الأرجحية لو كان التتابع من عدة نيوكلوتيدات، هنا لابد من تضمين النموذج جزء النهج process حيث أن الإستبدالات بين القواعد ستلعب دورا، والذى ستمثله المصفوفة التالية:

$$P = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

حيث أن ترتيب القواعد فى المصفوفة سيكون أبجديا أى A ثم C ثم G ثم T، ومن تلك القيم فإن احتمال بقاء A دون تغير هو 0.976 وان تستبدل بـ C هو 0.01 أو تستبدل بـ G هو 0.007 وهكذا لباقى الإحتمالات السته عشر. وجزء النموذج الخاص بالمكون composition والتى سنرمز لها بـ π، دعنا نقول أن نسبهم ستكون $\pi = [0.1, 0.4, 0.2, 0.3]$

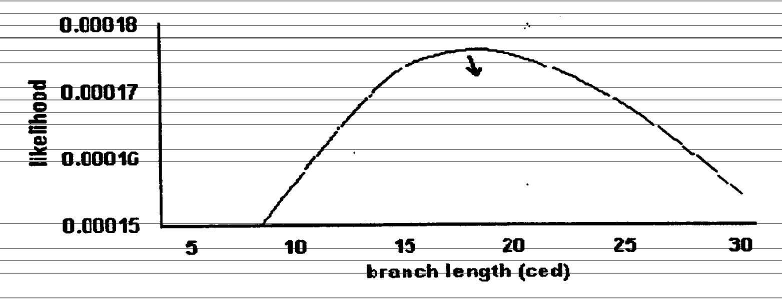
وبفرض وجود تتابعان هما ss و CCGT فسوف يربطهما فرع واحد بسيط (الشجرة)، إذن أرجحية هذا الفرع يمكن أن تحسب من العلاقة التالية:

$$= \pi_c P_{c-c} \; \pi_c P_{c-c} \; \pi_a P_{a-g} \; \pi_t P_{t-t} = 0.4 \times 0.983 \times 0.4 \times 0.983$$
$$\times 0.1 \times 0.007 \times 0.3 \times 0.79 = 0.0000300$$

وحيث أن فرع الشجرة السابق بسيط، فيجب حساب فروع الشجرة عندما تكون متغيرة الطول (كما هو الحال فى معظم الأحيان) وذلك نسبة لوحدات زمنية تطورية (للسهولة إفترض أنها وحدات بعد تطورى إعتبارية certain evolutionary distance [ced]، وهى فى مثالنا السابق قدرها وحدة واحدة. ولكن ما التغيرات المنتظرة لمصفوفات النهج إذا ما كانت قيم ced أكبر من واحد (تضرب المصفوفة فى نفسها بعدد قيم وحدات الزمن التطورى)، كما فى المصفوفات التالية:

$$P^2 = \begin{bmatrix} 0.953 & 0.02 & 0.013 & 0.015 \\ 0.005 & 0.966 & 0.01 & 0.02 \\ 0.007 & 0.02 & 0.959 & 0.015 \\ 0.005 & 0.026 & 0.01 & 0.959 \end{bmatrix} \;\&\; P^3 = \begin{bmatrix} 0.93 & 0.029 & 0.019 & 0.022 \\ 0.007 & 0.949 & 0.015 & 0.029 \\ 0.01 & 0.029 & 0.939 & 0.022 \\ 0.007 & 0.038 & 0.015 & 0.94 \end{bmatrix}$$

ويمكننا ملاحظة أن زيادة طول الفرع يؤدى بالضرورة إلى خفض إحتمالات أن تبقى القواعد دون تغير (الصف القطرى on diagonal) بينما إحتمالات الإستبدالات تزيد (الصفوف غير القطرية off diagonal). والرسم البيانى المبين بشكل (٤-٧) يوضح العلاقة بين وحدات البعد التطورى وقيم الأرجحية.



شكل (٤-٧) : العلاقة البيانية بين وحدات البعد التطورى وقيم الأرجحية المحسوبة.

ويتضح أن قيم الأرجحية تصل إلى حدها الأقصى عند وحدات بعد تطورى تتراوح بين ١٥ — ٢٠ ced ، كذلك إذا مازادت قيم البعد بقدر كبير فإن مصفوفة النهج ستؤول إلى قيم المكون π ، كما هو موضح بالمصفوفة التالية:

$$P^{10^6} = \begin{bmatrix} 0.1 & 0.4 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.3 \end{bmatrix}$$

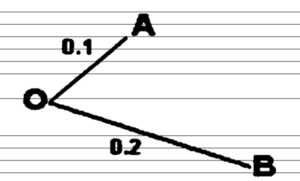
وعادة تستعمل المصفوفات السابقة فى صورة لوغارتمية logarithmic لصغر الأرقام ولمحاولة فصل المنهج عن المكون فى النموذج، كما أنها تتيح لنا حساب طول الفرع نسبة إلى معدل الاستبدال لكل موقع substitution per site خلافا للوحدات الإعتبارية السابقة (ced) كذلك يتيح لنا التعامل مع أطوال للفروع من الصفر حتى مالانهاية، لذلك تصبح المصفوفة كالآتى:

$$\log P = \begin{bmatrix} -0.0244 & 0.0101 & 0.0067 & 0.0076 \\ 0.0025 & -0.0176 & 0.005 & 0.0101 \\ 0.0034 & 0.0101 & -0.021 & 0.0076 \\ 0.0025 & 0.0134 & 0.005 & -0.021 \end{bmatrix}$$

حيث يلاحظ ان مجموع الصفوف سيؤول للصفر فى هذه الحالة. ويجب تحويل المصفوفة للأساس الأسى حتى تعبر عن احتمالات منسوبة للإستبدالات لكل موقع كالآتى:

$$scaled\ P = \begin{bmatrix} -0.122 & 0.05 & 0.034 & 0.038 \\ 0.05 & -0.353 & 0.101 & 0.202 \\ 0.034 & 0.101 & -0.21 & 0.076 \\ 0.038 & 0.202 & 0.076 & -0.315 \end{bmatrix}$$

ولحساب العلاقة بين تتابعين مختلفين (A & B) عند أطوال أفرع مختلفة، كما هو مبين كالتالى:



حيث تتابع A هو CCAT وتتابع B هو CCGT و تم ربطهما بمنشأ غير معروف (O)، وبإستعمال المصفوفات لمعدلات إستبدال قدرها $P_{0.1}, P_{0.2}, P_{0.3}$ يمكن حساب الأرجحية بعدة طرق، ابسطها : تعتمد على ان المسافة الكلية بينهما هى 0.3(من A إلى O ومن O إلى B)إذن تستعمل احتمالات المصفوفة $P_{0.3}$، ومنها نحصل على التقدير التالى:

$$\pi_c P_{c\text{-}c}\ \pi_c P_{c\text{-}c}\ \pi_a P_{a\text{-}g}\ \pi_t P_{t\text{-}t}$$

$$= 0.4 \times 0.786 \times 0.4 \times 0.786 \times .1 \times 0.08 \times 0.3 \times 0.747 = 0.000177$$

٢.٧. برامج الكمبيوتر للعلاقات الفيلوجينية.

Software for phylogenetic relations.

حاولنا فى الأجزاء السابقة التعريف بالأسس الحسابية والإحصائية لطرق تحديد العلاقات الفيلوجينية لعينات من تتابعات النيوكلوتيدات أو البروتينات لتفهم الأسس التى بنيت عليها تلك الطرق مع بيان مميزات وعيوب كل منها. ومما سبق يتضح أن هذه الطرق تعتمد على أساليب رياضية معقدة وصعبة وأن تنفيذها يدويا يعتبر متعب للغاية ويحتاج لزمن طويل، لذلك فإن تلك الحسابات تتم إعتمادا على برامج عديدة للكمبيوتر لتسهيل المهمة وتوفير الوقت. وخلال السنوات القليلة الماضية توفر عدد كبير من تلك البرامج المتخصصة وصل عددها الآن لأكثر من ٣٠٠ برنامج مختلف تتباين فيما بينها من حيث طبيعة اللغات والأنظمة المستخدمة، ومنها ماهو متاح للجميع بدون مقابل (خصوصا الأصدارات القديمة منها) وأغلبها يباع لمن يرغب. وعادة يمكن تبويب تلك البرمجيات تبعا للغرض الذى صممت من اجله، ويمكن تلخيص بعض منها كالتالى:

PHYLIP -1

وهو من أشهر البرامج فى الدراسات الفيلوجينية، ويمكن الحصول عليه مجانا من الـ internet من الموقع

http://evolution.gs.washington.edu/phylip.html

وهو متوافق مع معظم الأنظمة العالمية للكمبيوتر، وعن طريقه يمكن القيام بحسابات الطرق المقتصدة parsimony ومصفوفات البعد distance matrix والأرجحية العظمى maximum likelihood وغيرها مستعملا بيانات من تتابعات الـ DNA أو الـ RNA أو البروتينات أو مواقع القصر أو أى بيانات غير مستمرة (1/0) أوالتكرارات الجينية. وهو (مع البرنامج التالى) يمثل أكثر من ٨٠٪ من بيانات القرابة الفيلوجينية المنشورة فى الدوريات العالمية المتخصصة.

PAUP -2

وهى حزمة من البرامج وتعنى Phylogenetic Analysis Using Parsimony حيث صمم فى الأصل لهذا الغرض ثم طور فيما بعد ليقوم بكافة الحسابات التى يقوم بها البرنامج السابق، وهو ليس بالمجان وإن كان ثمنه غير باهظ (حوالى ٩٠ دولار لنظام Windows). ولمزيد من التفاصيل عن البرنامج يمكن اللجوء إلى الموقع:

. http://paup.csit.fsu.edu/

MacClade -3

يعتبر من البرامج الرائدة فى تحليل العلاقات التطورية مستعملا عدد متباين من طرز البيانات منها الصفات المظهرية غير المستمرة والبيانات الجزيئية، ويباع بحوالى ١٢٥ دولارا، ولمزيد من التفاصيل يمكن الأطلاع على الموقع:

http://phylogeny.arizona.edu/macclade/macclade.html

Hennig86 -4

وهو برنامج سريع لتقدير المعايير المقتصدة وطرق للبحث عن أنسب التفرعات للشجرات الفيلوجينية، والبرنامج مشفر وللحصول على الشفرة لابد من دفع إشتراك ومصاريف البريد. والبرنامج يمكنه التعامل مع حوالى ١٨٠ عينة taxa وحوالى ٩٩٩ صفة مختلفة، ولمزيد من التفاصيل عنه يمكن الرجوع إلى:

Farris, J.S. 1989, Hennig86: a PC-DOS program for phylogenetic analysis. *Cladistics* 5: 163.

Random Cladistics -5

وهو برنامج للشجرات يعتمد فى تقديراته على برنامج Hennig86 وهوقادر عن البحث عن الجزر بين الشجرات، والوصف المفصل للبرنامج موجود على الموقع:

http://research.amnh.org/~siddall/rc.html

حيث يمكن الحصول على نسخة منه بلغة الـ DOS.

AutoDecay -6

وهو برنامج خاص يمكنه أن يبنى معاملات التداعيات decay indices اعتمادا على الشجرات المبنية على أحد البرامج الأخرى وهو برنامج PAUP. وهو مكتوب بلغة Perl ومتوافق مع أغلب أنظمة الكمبيوتر المتاحة، ويمكن الحصول عليه من شبكة المعلومات من الموقع:

http://www.bergianska.se/index_forskning_soft.html.

DNA Stacks -7

وهو برنامج لتحليل تتابعات الـ DNA ورصها ولا يقوم ببناء العلاقات الفيلوجينية، وهو مساعد للبرامج الأخرى مثل PAUP و PHYLIP ولكنه صالح لنظام Macintosh فقط. ويمكن الحصول عليه من شبكة المعلومات من الموقع:

http://biology.fullerton.edu/deernisse/dnastacks.html.

TreeRot -8

وهو برنامج لعمل الشجرات إعتمادا على الإحصائيات المقتصدة ولكن متوافق مع نظام Macintosh ولمزيد من التفاصيل يمكن الرجوع إلى الموقع:

.http://people.bu.edu/msoren/TreeRot.html

RA -9

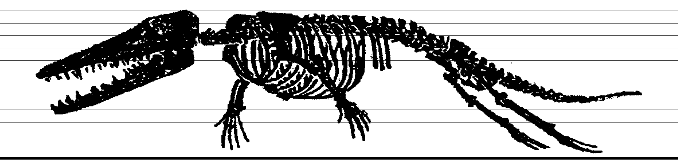
وهو برنامج صغير لتحليل بيانات من تتابعات النيوكلوتيداتوهو متوافق مع العديد من الأنظمة ويعتبر رخيص الثمن (حوالى ٢٠ دولار).

٧.٤. تطبيقات الفيلوجينيا Applications of phylogenetics.

الفيلوجينيا هى محاولات لدراسة التاريخ التطورى لمجموعة من الكائنات تمثل وحدات تقسيمية taxa ما على أسس وراثية. وحديثا تركزت هذه الدراسات على فحص التتابعات الجينية أو البروتينية (الجزيئات الحيوية) لذلك قد تعرف أيضا بإسم الجينومكس المقارن comparative genomics. والفيلوجينيا تختص بدراسة وتفهم موضوعين رئيسيين ، أولهما بناء الشجرات التطورية مع تحديد تفرعاتها المختلفة ، وثانيهما هو إستغلال تلك الشجرات لمعرفة مسارات تطورية جديدة، كذلك لا يجب أن نغفل الهدف التقليدى للفيلوجينيا ألا وهو التقسيم classification & systematics. وفيما يلى بعض من أهم تطبيقات تلك الدراسات.

٧.٤.١. التطور Evolution.

تطور الكائنات موضوع واسع وكبير، لكن من الأمثلة الشيقة فى هذا المجال هى الدراسات المستفيضة التى أجريت لمعرفة أسلاف الحيتان whale ancestors . فقد ظل موضوع تطور الحيتان والدرافيل (cetaceans) من المعضلات غير المفهومة فى علم التطور لزمن طويل، فكيف يمكن تفسير أن هذا الحيوان الثديي mammal الكبير قد إرتد إلى البحر مرة أخرى وتأقلم على العيش فى تلك البيئة المائية على عكس مسار التطور المعروف بأن الكائنات البرية terrestrial قد خرجت من البحر منذ ملايين السنين خلال مسارها التطورى ! لكن حديثا وفى أواخر القرن الماضى (١٩٨٠ — ١٩٩٥) تم إكتشاف عدد من الحفريات التى تمثل حلقات الوصل بين الثدييات الحافرية artiodactylans والحيتان البحرية، فقد اكتشف Gingerich أول تلك الحفريات فى باكستان وسميت *Packcetus* ثم حفريات الحوت ذو الأرجل *Ambulocetus* فى باكستان ومصر بواسطة Thewissen، وشكل (٧-٥) يبين هذا الحيوان المنقرض.



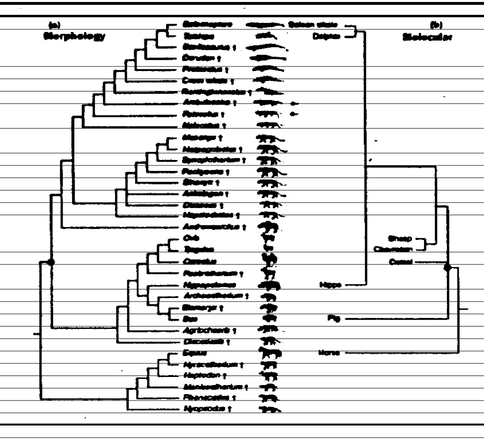
شكل (٧-٥) : حفرية الثدى المنقرض من حوالى ٤٠ مليون عام المعروف بأسم *Ambulocetus*.

وببناء الشجرة الفيلوجينية للثديات الحافرية إعتمادا على الصفات المظهرية لما هو باق منها ولمجموعة الحفريات fossils لحيوانات تابعة لها ولكنها إنقرضت، يتضح أن الحيتان والدرافيل قد نشأت من سلف مشترك (ثديي حافرى) يمثل حلقة الوصل بين الحيوانات البرية والمائية يعرف بأسم *Ambulocetus*. ولكن عندما بنيت الشجرة بأستعمال البيانات الجزيئية من تتابعات الـ DNA المستخلص من الميتوكوندريا والأنوية mt DNA & nuclear DNA لعدد من ذات الحافر، إتضح أن الحوت أكثر قرابة إلى فرس النهر (سيد قشطة) ويمثلان فصيلتان شقيقتان sister clade، وأن فرس النهر أقرب للحوت أكثر من قرابته لباقى ذات الحافر مثل الأبقار و الحصان وغيرها، كما هو موضح بشكل (٧-٦).

٧. ٤. ٢. أدلة الطب الشرعي Forensic medicine evidence.

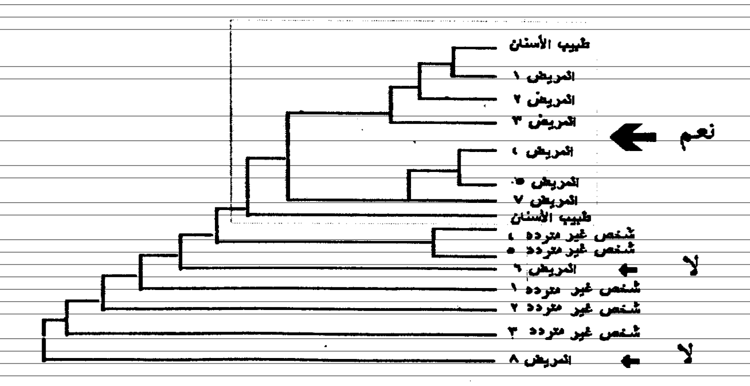
أصبح استعمال تقنيات البيولوجيا الجزيئية وخصوصا تكنيك البصمة الوراثية DNA fingerprinting فى مجالات الطب الشرعي فى قضايا البنوة وإثبات النسب وكذلك للتعرف على المتهمين فى الجنايات من الأمور شائعة الاستعمال منذ حوالي ربع قرن مضى من الزمان. ولكن ما يهمنا فى هنا هو أمكانية استعمال الشجرات الفيلوجينية المعتمدة على تحليل تتابعات النيوكلوتيدات فى مجالات الطب الشرعي أو الطب الوقائي.

ففى دراسة شيقة قام بها كل من Ou وزملاؤه سنة ١٩٩٢ لتحديد مصدر الإصابة بفيروس مرض الإيدز بين مجموعة من المرضى المترددين على عيادة أحد أطباء الأسنان بالولايات المتحدة، قاما بعزل تلك الفيروسات من هؤلاء المرضى وطبيب الأسنان ومن عدد أخر من الأفراد المصابين بالإيدز ولكن غير مترددين على عيادة هذا الطبيب (عينة ضابطة)، وقاما



شكل (٧-٦) : شجرات فيلوجينية للثدييات ذات الحافر (a) بناء على الصفات المظهرية morphology والحفريات ، (b) بناء على تتابعات الـ DNA الميتوكوندرى والنووى (molecular). حيوانات منقرضة = †.

ببناء شجرة فيلوجينية إعتمادا على تحليل تتابعات النيوكلوتيدات لتلك العزلات من الفيروسات، كما هو موضح بشكل (٧-٧). فأمكنهم تحديد المرضى الذين اصيبوا بالمرض عن طريق عيادة هذا الطبيب والمرضى الذين أصيبوا من مصادر أخرى.



شكل (٧-٧) : شجرة فيلوجينية لتتابعات فصائل فيروس الأيدز.

٧. ٤. ٣. التنبؤ بوظائف الجينات Gene function prediction.

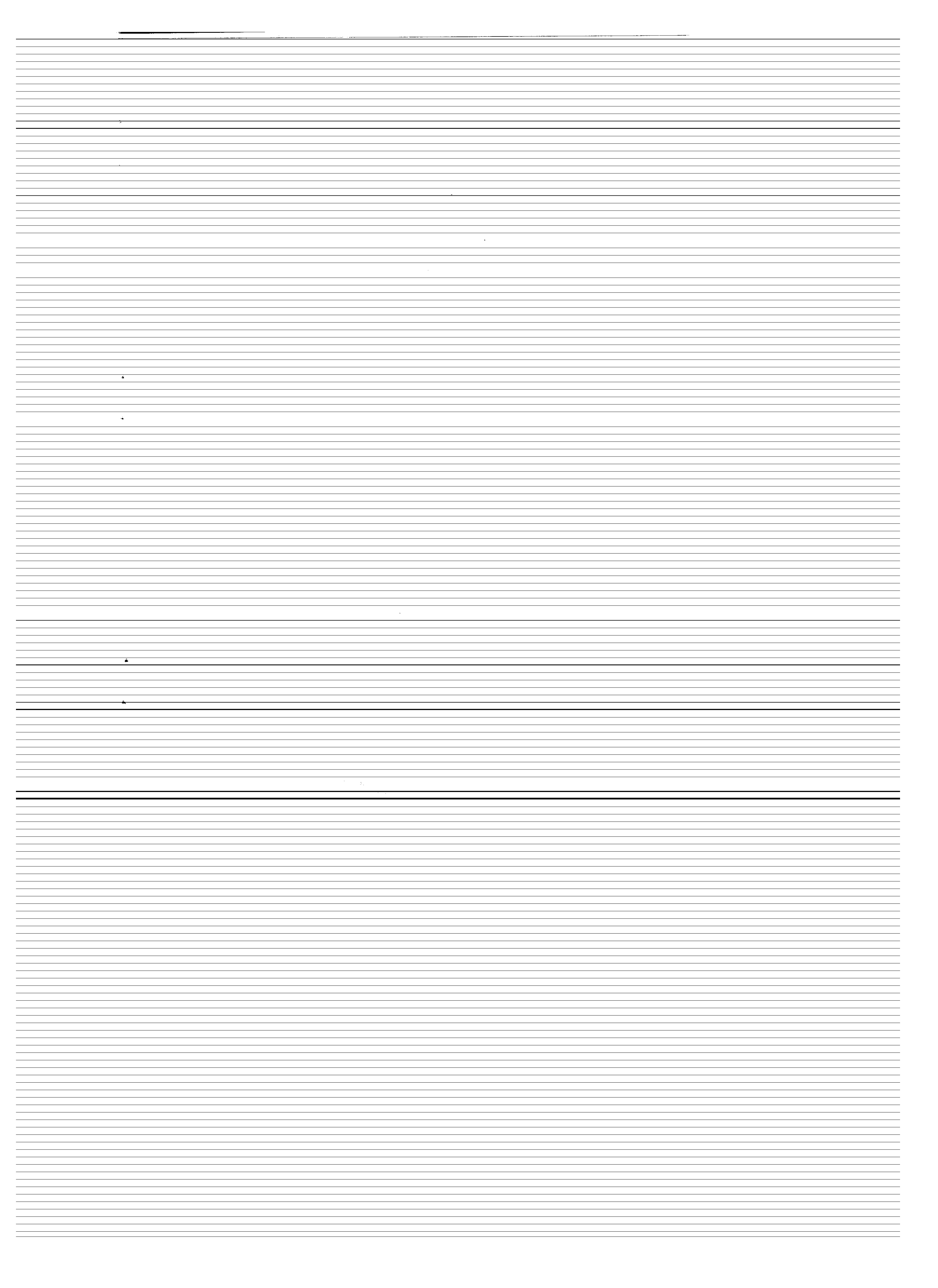
خلال السنوات القليلة الماضية ظهرت الحاجة الملحة للتعرف على وظائف الجينات وتحديدها وذلك مواكبتا لازدهار دراسات الجينومكس فى العديد من الكائنات، وخلال تلك الدراسات تم عزل تتابعات (جينات) لم تكن معروفة من قبل وكان على الوراثيون أن يتعرفوا أو على الأقل أن يحاولوا التنبؤ بوظائفها. وفى سبيل ذلك إتبع العلماء طرق عديدة ولكن اكثرها شيوعا هو محاولة دراسة درجات التشابه similarity بين تلك الجينات مجهولة الوظيفة وشبيهاتها (بناء على صف alignment تتابعات النيوكلوتيدات – الباب السادس) من الجينات معلومة الوظيفة ثم محاولة التنبؤ بوظيفة هذا الجين مجهول الوظيفة. وهذه الطرق أثبتت نجاحا لا بأس به فى التعرف على وظائف العديد من الجينات، ولكن فى بعض الحالات فأن التشابه الجزيئى لم يعكس بالضرورة التشابه الوظيفى، يمكن الرجوع إلى Hillis سنة ١٩٩٤ لتحديد بعض من تلك الحالات النادرة. لذلك يفضل العديد من الوراثيين إستعمال اكثر من طريقة لهذا الفرض وعدم الاكتفاء بدراسة التشابه الجزيئى. خصوصا أن مفهوم التماثل الوراثى homology غير محدد، فعلى سبيل المثال الجدول التالى يلخص عدد من مصطلحات التماثل التى قد تتداخل فى مفاهيمها مع بعضها البعض:

المصطلح	المعنى
Homolog	الجينات المتشابهة التى يربطها سلف واحد مشترك (مثل كل جينات الجلوبين globins فى الثدييات).
Ortholog	الجينات المتشابهة التى إنفصلت عن بعضها diverged بعد حادثة (خطوة) تطورية ما speciation event (مثل جينيβ-globin فى الإنسان والفأر).
Paralog	الجينات المتشابهة التى إنفصلت عن بعضها diverged بعد حدوث تضاعف جينىgene duplication (مثل جيني β-globin و γ-globin فى الإنسان).
Xenolog	الجينات المتشابهة التى إنفصلت عن بعضها البعض diverged بعد حدوث إنتقال جزئى لأحد الجينات lateral gene transfer (مثل جينات مقاومة المضادات الحيوية فى البكتيريا).

ويعتقد العلماء أن طرق الصف لتحديد التماثل غير قادرة على عكس البعد التطورى لهذه الجينات لذلك اقترح Eisen سنة ١٩٩٨ طريقة جديدة تجمع بين المعلومات الفيلوجينية (الشجرات) وطرق الصف التقليدية للتنبؤ بوظائف الجينات والتى أطلق عليها أسم "الفيلوجينومكس" Phylogenomics. وشكل (٧- ٨) يلخص أهم خطوات هذه الطريقة.

1 حدد الجين غير معلوم الوظيفة	5
2 تعرف على مجموعة الجينات المناظرة	3 4 5 1 2 6
3 بناء على صف التتابعات alignment حدد الشجرة الفيلوجينيه لهذه المجموعه من الجينات	1 2 3 4 5 6
4 وقع على الشجرة وظائف الجينات المعروفة ما أمكن حيث الجينات ١ و ٢ حدد لها وظيفة واحدة والجينات ٤ و ٦ حدد لها وظيفة واحدة أخرى بينما الجين ٣ فغير معلوم الوظيفة مثله مثل الجين تحت الدراسة رقم ٥	? ? 1 2 3 4 5 6
5 يمكن التنبؤ بأن وظيفة الجين ٥ قد تكون مماثلة للجينات ٤ و ٦ بينما الجين ٣ فوظيفته اقرب للجينات ١ و ٢	1 2 3 4 5 6

ومع افضلية هذه الطريقة ونجاحها فى التعرف على وظائف العديد من الجينات مجهولة الوظيفة إلا أنها غير مؤكدة بالقدر الكافى فمازال هناك حالات لا يمكن التعرف عليها لذلك وجب البحث على طرق أكثر دقة وتحديدا.



٨. البروتيوميكس

Proteomics

إعداد : أحمد الشهاوى

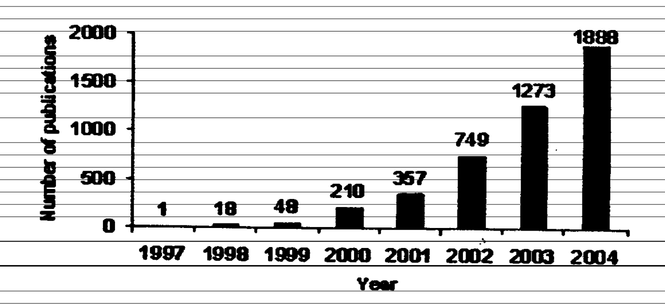
١.٨. مقدمه.

بعد الإنتهاء من مشاريع قراءة التتابع الكامل لجينوم العديد من الكائنات الحيه مثل جينوم نبات الأرابيدوبسيس Arabidopsis ، جينوم الإنسان Human genome ، ....الخ، بدأ إهتمام العلماء بعصر مابعد الجينوم post-genomic era. هذا يعنى كيفية الإستفاده من المعلومات الجينوميه الهائله التى تم تجميعها وتخزينها لفهم الوظائف والأنظمه الخلويه والإختلافات بينها. ففى الوقت الحالى توجد فجوه كبيره فى الوسائل والطرق لربط المعلومات الهائله المتاحه عن التراكيب الجينوميه المختلفه ووظائف البروتينات المشفره من هذه المعلومات الوراثيه (ربط تركيب الجينوم بوظيفته).

بالنظر إلى بعض الجينومات التى تم الإنتهاء من تحديد تتابعها يمكن توقع التحديات الضخمه التى تنتظر العلماء فى عصر مابعد الجينوم. فعلى سبيل المثال، فإن جينوم خميرة الخباز (*Saccharomyces serevisiae*) والذى تم الإنتهاء من قراءته، فإن حوالى ٣٠٪ من الـ Open Reading Frames (ORFs) مازال غير معروف وظيفتها وتحتاج إلى طرق تعتمد على تحليل ودراسه الجينوم ككل.إن تركيب وتنظيم الجينومات الأكثر تعقيدا مثل جينوم الإنسان تجعل المهمه أكثر صعوبه مقارنتا بجينوم الخميره. وعند الأخذ فى الإعتبار تنظيم التعبير الجينى على مستوى النسخ، وما بعد النسخ، والترجمه، وما بعد الترجمه وكذلك مرونه الجينوم وقابليته للتغير فى حدود معينه، فإن المشكله تتضخم لدرجه لايمكن تصورها ويصبح تصورودراسه وفهم وظيفه الجينوم مهما كان صغيرا فى حجمه مهمه شبه مستحيله. وتصبح كل المحاولات لدراسه وظيفه الجينوم هى محاولات لفهم جزئى لوظيفه الجينوم. لذلك يحتاج عصر مابعد الجينوم لطرق تحليل الجينوم بالكامل whole system-based analysis لمحاوله الفهم لوظيفه الجينوم ككل. ومن أهم الوسائل المتاحه لعصر مابعد الجينوم هو التحليل الكمى لتعبير

بعض الجينات على مستوى النسخ quantitative monitoring of expression level لعدد كبير من الجينات معا من خلال إستخدام microarray technology. كذلك يمكن دراسه وظيفه جين معين من خلال دراسه التغاير فى التعبير الجينى differential gene expression، ولكن فى حاله الكائنات المعقده فإن المعلومات التى يتم الحصول عليها من دراسات التعبير الجينى على مستوى النسخ لا تعكس التغيرات التى تحدث على مستوى البروتين فى معظم الأحيان. هذا يرجع إلى طرق التنظيم المختلفه والعديده التى تحدث أثناء الترجمه ومابعد الترجمه translational and posttranslational regulations. فقد أشارت دراسات عديده على ضعف العلاقه بين تعبير الـ RNA والبروتين، وبالتالى فمن الصعوبه إستخدام مستوى النسخ (RNA) لتوقع مستوى الترجمه (البروتين)، وتحورات مابعد الترجمه post-translational processing. ففى دراسه على جينوم الخميره، دلت النتائج على انه لابد من وجود طرق أدق لدراسة التعبير الجينى للجينوم ككل. وقد شملت هذه الدراسه التعليم بواسطة العوامل الوراثيه المتنقله transposon tagging لعدد ٣١ جين من جينات الإنقسام الميوزى meiosis لدراسه تعبيرهم الجينى على مستوى البروتين، لكن لم يتمكنوا إلا من تحديد وظيفة ١٧ جين فقط من هذه الجينات عند دراستها بتكنيك الـ microarray. من ذلك يمكن القول بإن العلاقه بين التعبير الجينى على مستوى الـ RNA، والتعبير الجينى على مستوى البروتين ليست فقط علاقه معقده ولكنها قد تكون مضلله.

لايوجد فى الوقت الحالى طرق عامه لتحليل وفهم الوظائف البيولوجيه للجينات. هذه الطرق يجب أن تتضمن العديد من طرق تحليل الجزيئات البيولوجيه المختلفه المشتركه فى الوظائف الخلويه فى نفس الوقت. تعتبر البروتينات من أهم الجزيئات البيولوجيه التى تقوم بالوظائف الخلويه بطريقه مباشره.ولقد تم إستخدام مصطلح البروتيومكس proteomics بواسطه Wilkins *et al.* سنة ١٩٩٦ لوصف دراسه كل البروتينات المنتجه بواسطة الجينوم ككل وفى صوره كميه عند وقت معين أو إستجابه لظروف معينه. هذا المجال مازال فى مراحله الأوليه ويحتاج إلى جهد كبير لتطويره فى المستقبل، و بالرغم من أن هذا المجال مازال فى مراحل تطوره الأولى فإنه يمثل نقطة جذب للباحثين كما هو موضح فى الشكل التالى لعدد الأبحاث المنشوره والتى تحتوى على كلمه بروتيومكس (شكل ٨-١).



شكل(٨-١): عدد الأبحاث المنشوره فى الفتره من ١٩٩٧ حتى ٢٠٠٤. تم الحصول على هذه الأعداد بإجراء بحث فى الـ PubMed بإستخدام كلمه proteomics.

٨. ٢. عزل وفصل البروتينات.

بالرغم من أن مجال البروتيومكس proteomics حديث نسبيا فإن بعض تقنياته مستخدمة من حوالى ٢٥ سنه. مثل الفصل الكهربى للبروتينات فى إتجاهين 2D gel electrophoresis الذى يستخدم لفصل ودراسة البروتينات منذ أكثر من ٢٥ سنه. كذلك فإن إستخدام Mass Spectroscopy مع الفصل الكهربى للبروتينات فى إتجاهين 2D gel electrophoresis كونا أول جيل من نظم دراسة البروتيومكس first generation of proteomic platforms. و أثناء السنوات القليله الماضيه تم تطوير العديد من الطرق لفصل وتنقية البروتينات والتى تعتبر الأساس للجيل الأول من نظم دراسه البروتيومكس وفيمايلى ملخص لهذه الطرق.

٨. ٢. ١. عزل البروتين Protein Isolation.

اول واهم خطوه فى دراسه البروتيوم proteome (المعين البروتينى) هى إستخلاص البروتين من البيئه الموجود بها. ففى هذه الخطوه يجب الإهتمام بعزل البروتيوم proteome فى اقرب صوره كان موجود عليها فى الخليه وتقليل تأثير طريقه الفصل. يجب التأكيد على أنه لاتوجد طريقه عامه لفصل البروتينات نظرا لإختلاف

طبيعة العينات المستخدمه ولكن توجد طرق مختلفه لفصل البروتينات بطريقه مرضيه من العينات المختلفه. فلفصل بروتينات السيتوسول cytosolic proteins يتم تحليل الخلايا بطريقه بسيطه وإستخدام الـ supernatant الناتج. هناك العديد من الطرق المختلفه لعزل البروتينات وهى متاحه على الموقع التالى بشبكة المعلومات الدوليه (http://expasy.cbr.nrc.ca/ch2d/protocols)، وبعض منها قد تكون بسيطه كإستخدام صدمه الضغط الأسموزى osmotic shock، أو طرق أعنف مثل إستخدام الموجات الفوق صوتيه ultrasonication. إن إختيار الطريقه المسخدمه لفصل البروتين تتوقف على السهولتها وملاءمتها لفصل البروتينات فى إتجاهين (تركيز الملح والمذيبات الأيونيه). وبالرغم من أن معظم هذه الطرق مدروسه بطريقه جيده إلا أن فصل بعض مكونات البروتيوم proteome مثل البروتينات الغير محبه للماء hydrophobic proteins مازال من الصعوبه بمكان. ففى السنوات الأخيره تم تطوير بعض الكيماويات والطرق لتحسين فصل البروتينات الغير محبه للماء hydrophobic proteins، ومع هذا التقدم النسبى فى طرق فصل البروتينات للتحليل ثنائى الإتجاه إلا أن نتائج دراسات البروتيوم تشير إلى أن الإختلافات الخلويه فى العينه المستخدمه وكذلك الإختلاف فى أماكن تخزين البروين داخل الخلايا يؤثر فى نتائج هذه الدراسات.

٨.٢.١.١. العينه المتجانسه Homogenous Sample.

الخلايا المنزرعه cell cultures ربما تكون من أفضل مصادر البروتين لدراسة البروتيوم حيث أنها تمثل مجموعه من الخلايا المتجانسه وتوفر بروتيوم متحكم فيه إلى حد ما. كذلك فإن ظروف نمو وتحلل الخلايا يمكن التحكم فيه والذى يقلل من الإختلافات بين المزارع والإختلافات غير الحقيقيه artifacts.

٨.٢.١.٢. العينه الغير متجانسه Heterogeneous Sample.

العينات غير المتجانسه مثل عينات الأنسجه تجعل تحليل البروتيوم لها عمليه معقده للغايه وذلك بسبب الإختلافات الكبيره فى طريقة الحصول على العينات وطريقة إستخلاص البروتين. فمثلا، عينات الأنسجه التى تستخدم فى دراسات البروتيوم فى الإنسان يتم تجميعها من مرضى من المستشفيات. حيث لاتوجد عملية لرقابه كافيه على جمع وحفظ هذه العينات والذى يمثل مصدر كبير للإختالافات غير الحقيقيه فى

البروتيوم نتيجة الإستجابه للمؤثرات الخارجيه. عينات الأنسجه غير متجانسه بطبيعتها (تحتوى على أنواع خلايا مختلفه بأحجام مختلفه وبنسب خلايا مختلفه) وهذا يؤدى إلى وجود إختلافات فى البروتيوم من نفس النسيج أو بين الأنسجه المختلفه. لذلك فإن النتائج المستخلصه من مثل هذه الدراسات يجب الوصول اليها بعد دراسة عدد كبير من العينات. فى الواقع فإن البروتيوم المستخلص من عينات الأنسجه يتكون من عديد من البروتيوم كل منها ناتج من مجموعه من الخلايا فى العينه أو النسيج. وفى هذه الحاله فإن البروتيوم المختلط الناتج لايمكن مقارنته حتى بدراسة عديد من العينات.

٨. ٢. ١. ٣. مجاميع الخلايا داخل العينه Sub-Population of Cells.

أصبح من الضرورى فصل عشيره من الخلايا من العينه البيولوجيه لتحليلها وتحديد المعلومات المستمده من كل بروتيوم. حديثا تم إستخدام أنواع متقدمه من أجهزه فصل الخلايا cell sorters وطرق مناعيه متقدمه لفصل تحت عشائر الخلايا لتحليل البروتيوم.

٨. ٢. ١. ٤. التراكيب تحت الخلويه Subcellular components.

يمكن أيضا تبسيط البروتيوم إلى مجموعه البروتينات المستخلصه من أحد مكونات الخليه حيث أن نقل وتوزيع البروتينات إلى عضيات معينه داخل الخليه عمليه معروفه فى مجال بيولوجيا الخليه cell biology. فيمكن فصل أحد عضيات الخليه والتعامل مع بروتينات هذه العضيه على أنها بروتيوم مستقل.

٨. ٢. ٢. الفصل الكهربى للبروتينات فى الإتجاهين.

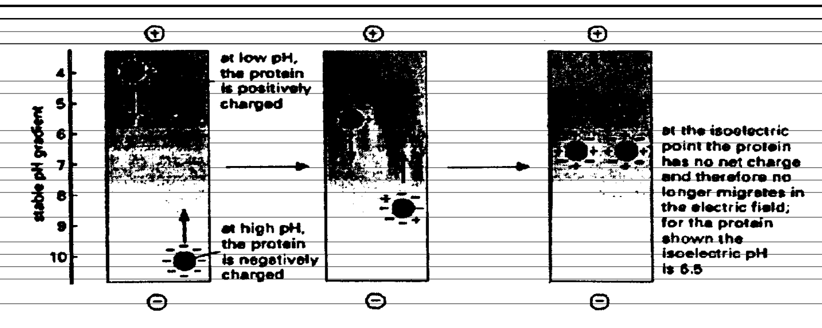
Separation by 2D electrophoresis.

بعد عزل وتنقية البروتينات يلزم فصلها بطريقه معينه لفصل مكونات مخلوط البروتين الى بروتينات فرديه. ففى مجال البروتيومكس، فإن طريقة الفصل المفضله هى الفصل الكهربى فى إتجاهين 2D electrophoresis. هذه الطريقه يمكن أن تستخدم لفصل عده الأف من البروتينات فى تجربه واحده.

٨. ٢. ١.٢. أساسيات فصل البروتينات فى إتجاهين.

Principles of 2D separation.

بدأ الإستخدام الأمثل للفصل الكهربى للبروتينات فى إتجاهين بعد تطور شرائط التدرج فى الـ pH الثابتـه Immobilized pH Gradient (IPG) فـى الثمانينـات مـن القـرن الماضى، والذى يـتم ببلمـره مجموعـه مـن الجزيئـات الأحاديـة monomers والتـى تحمـل وظيفه الـ ampholytes فى وجود acrylamide. لذلك بتغيير تركيز هـذه الجزيئـات على طول الشريط ينتج تدرج ثابت فى الـ pH. وتوجد تجاريا أشرطه تحتوى على تـدرج pH فى مدى مختلف لفصل البروتينات فى إتجاهين. الإتجاه الأول للفصل (تركيـز البروتينـات تجـاه نقطـه التعـادل الكهربـى) Isoelectric Foucsing. للفصـل فـى هـذا الإتجـاه، فـإن مستخلص البروتين يستخدم لتشبيع وإعاده تمدد شريط ال IPG قبل الفصل. هذا يضمن ان البروتين سيكون موزع بشكل متساوى على الشـريط ويمنـع ترسـيب البـروتين لزيـاده تركيزه فى مناطق معينه على الشريط. يتم بعد ذلك إمرار تيـار كهربـى تـدريجى خـلال الشريط. البروتينات المشحونه بشحنه موجبه (أى انها تقع فـى منطقـه مـن الشـريط ذات pH أقل من درجه التعادل الكهربى *pI* لها) فإنهـا تتحـرك ناحيـه القطـب السـالب cathode وتواجه بدرجات pH أعلى حتى تصل إلى درجات التعادل الكهربـى *pI* لهـا وتصـبح عنـدها متعادله وتتوقف عن الحركه. البروتينات المشحونه بشحنه سالبه (أى انها تقع فى منطقه من الشريط ذات pH أعلى من درجه التعادل الكهربى *pI* لها) فإنها تتحـرك ناحيـه القطـب الموجب anode وتواجه بـدرجات pH أقل حتى تصل إلى درجـات التعـادل الكهربـى *pI* لهـا وتصبح عندها متعادله وتتوقف عن الحركـه. النتيجـه النهائيـه للفصـل، فـإن البروتينـات التى لها نفس نقطه التعادل الكهربى فإنها تركز وتوجه بإستمرار إلى منطقه من الشـريط ذات درجه pH مماثله لدرجة تعادلها الكهربى *pI* (شكل ٨-٢).



شكل(٨-٢) : فصل البروتين بالتركيز تجاه نقطه التعادل الكهربى. عند درجات الـ pH المنخفضه (تركيز مرتفع لأيونات الهيدروجين)، فإن البروتين يحمل شحنه موجبه.

الإتجاه الثانى للفصل، هو فصل البروتينات عل أساس وزنها الجزيئى SDS-PAGE، حيث يتم صب وبلمره محلول الأكريلاميد acrylamide بين قطعتين من الزجاج يفصل بينهما فاصل يتراوح سمكه من ١-١,٥ مم. يتم وضع الشريط الناتج من الفصل فى الإتجاه الأول أفقيا على SDS-PAGE (SDS-Polyacrylamide Gel Electrophoresis). يتم مرور تيار كهربى ويتم فصل البروتينات بناء على وزنها الجزيئى. أنظمة فصل البروتينات فى الإتجاه الأول والثانى متوفره تجاريا من مصادر مختلفه.

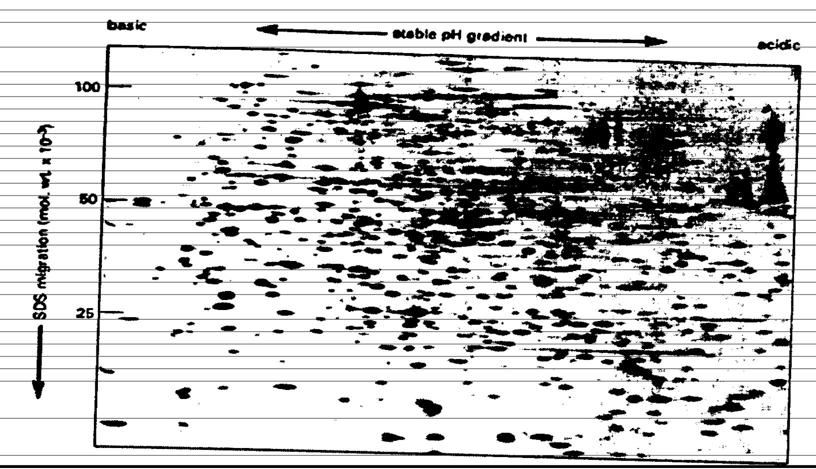
٨. ٢. ٢. ٢. تحديد وإظهار البروتينات المفصوله فى إتجاهين.

Resolution of Separated Proteins.

توجد طرق مختلفه لتحديد وإظهار البروتينات المفصوله فى إتجاهين. هذه الأنظمه تعتمد على تعليم البروتينات قبل أو بعد الفصل أو ربما بعد الفصل فى الإتجاه الأول. وأهم هذه الطرق هى: (١) تعليم البروتينات قبل الفصل Pre-Labeling of proteins ويندرج تحتها التعليم بالعناصر المشعه Radiolabeling، حيث يتم تعليم البروتينات قبل الفصل فى الإتجاه الأول عادة داخل الخليه *In vivo* بإستخدام العناصر المشعه radioisotopes. هذا التعليم يتم بإستخدام جزئ معلم (حامض امينى) مثل ^{35}S-methionine فى مزارع الخلايا، حيث يتم إظهار البروتينات المعلمه بعد ذلك بتعريض الجيل لفيلم X-ray أو شاشة phosphor imager كما فى

شكل(٨- ٣). هذا النوع من التعليم غير مناسب لدراسة عينات الحيوان والإنسان. ويندرج تحتها أيضا طرق التعليم الفلورسنتى Fluorescent labeling حيث يتم التعليم بصبغات فلورسنتيه متعادله كهربيا والتى ترتبط بالبروتين قبل الفصل الكهربى فى الإتجاه لأول. هذه الصبغات يعاب عليها أن حساسيتها أقل ولكن فى الفترة الأخيره تم تطوير صبغات يصل حد إيضاحها مستوى النانوجرام من البروتينات. إن كل هذه الصبغات الفلورسنتيه تحتاج إلى نظام خاص لتحديد شدة الفلورسنس fluorescence لمكان البروتين على الجيل.

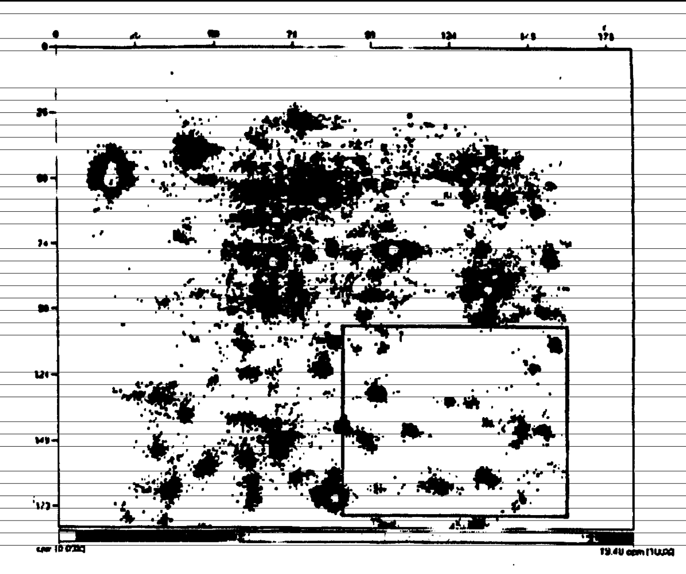
توجد طرق اخرى تسمى multiphoton detection (MPD) وتعتمد على تعليم البروتين *In vivo* بواسطه كميات قليله من اليود المشع I^{125} او I^{131}. هذا النظام يمكنه إظهار كميه من البروتين على مستوى الأوتومول (amol) attomol. فى هذه الحاله يتم إستخدام كميه من البروتين فى مستوى النانوجرام. شكل (٨- ٤) يوضح الفصل فى إتجاهين حوالى ١٥٠ نانوجرام من مستخلص بروتين بكتريا *E. coli* والتى تم تعليمها قبل الفصل بواسطة اليود المشع I^{125}. اما مجموعة طرق التعليم الثانية فتتميز بكونها طرق لتعليم البروتينات بعد الفصل Post-Labeling of Proteins، فتعليم البروتينات بعد الفصل هى الطريقه الشائعه لإظهار البروتينات بعد الفصل الكهربى فى إتجاهين ويتم ذلك بواسطه صبغات مثل Coomassie blue (حدود الإظهار بهذه الطريقه (LOD) Limit of Detection هى ٥٠ نانوجرام من البروتين) أو الصبغ بالفضه Silver Staining (حدود الصبغ بهذه الطريقه حوالى ٥ نانوجرام). توجد طرق أخرى لإظهار البروتينات معتمده على الصبغ العكسى reverse staining. بالرغم من أن الصبغ بالفضه Silver Staining يحتاج إلى عمل اكثر إلا انه يعتبر من أكفأ التكنيكات وأكثرها حساسيه لإظهار البروتينات على الجيل (فى الجيل). شكل (٨- ٥) يوضح جيل gel مصبوغ بالفضه مفصول فى إتجاهين لحوالى ١٥٠ ميكروجرام بروتين. ويلاحظ أن كل منطقه مصبوغه spot تحتوى على بروتين أو أكثر ويمكن إستخدامها لتحليل ودراسة هذه العينات.



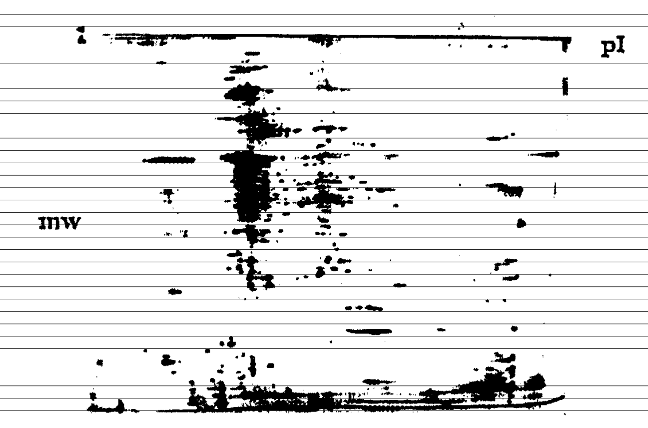
شكل (٨-٢): الفصل الكهربى للبروتين الكلى لـ *E. coli* فى إتجاهين معلمه بمخلوط من الأحماض الأمينيه المشعه قبل الفصل. الإتجاه الأفقى يوضح الفصل على حسب نقطة التعادل الكهربى *pI* والإتجاه الراسى يمثل الفصل على حسب الوزن الجزيئى.

كذلك توجد تكنيكات أخرى أكثر حساسيه تعتمد على التعليم الفلورسنتى.

هذه الطرق لا تظهر البروتينات فى صوره مرئيه على الجيل gel ولكنها تعتمد على اجهزه تستخلص البروتينات من الجيل وتوصفها مباشره. هذا الفصل ثنائى الإتجاه يعتبر من أكفأ الطرق لفصل البروتينات والذى يمكننا من فصل أكثر من ١٠٠٠٠ بروتين على الجيل الواحد.



شكل (٨-٤): الفصل الكهربي لبروتينات بكتريا *E. coli* (١٠٠ نانوجرام) في إتجاهين مصبوغه قبل الفصل باليود المشع ^{125}I . الإتجاه الأفقى يوضح الفصل على حسب نقطة التعادل الكهربي *pI* والإتجاه الرأسي يمثل الفصل على حسب الوزن الجزيئي.



شكل (٨-٥): الفصل الكهربي لبروتينات *Sulfolobus salfataricus* (١٥٠ مليجرام) في إتجاهين مصبوغه بعد الفصل بالفضه Silver staining. الإتجاه الأفقى يوضح الفصل على حسب نقطة التعادل الكهربي *pI* والإتجاه الرأسي يمثل الفصل على حسب الوزن الجزيئي.

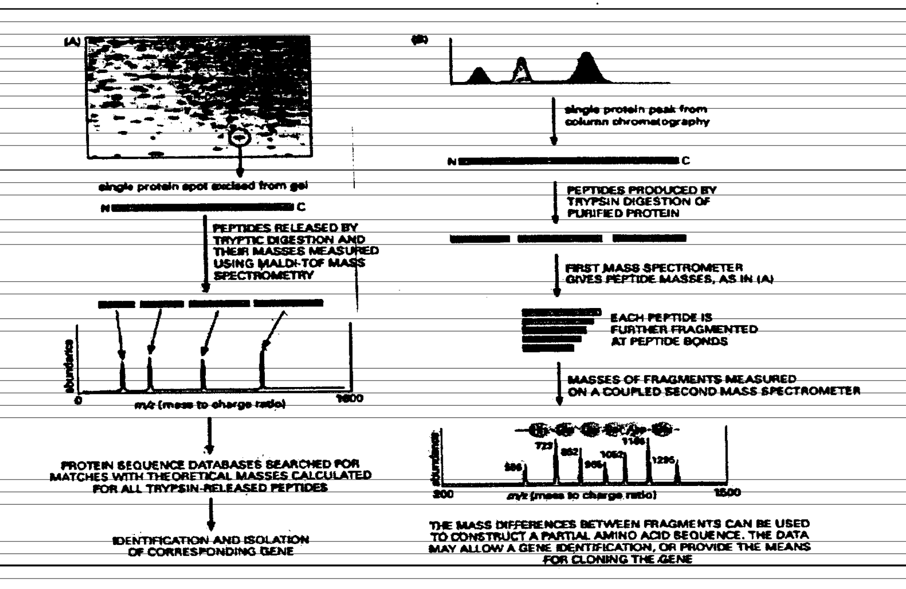
٨. ٣. التطورات الحديثه في دراسه البروتيومكس. NewDevelopments.

تشير الدراسات السابقه عن البروتيومكس أن مستوى إظهار وتحديد البروتينات المفصوله هو مستوى الفمتومول fmol. وهذا يعتبر مستوى مرتفع بالنسبه للبروتينات ذات التركيز المنخفض. وتعتبر هذه من المشاكل الصعبه بالنسبه للبروتينات منخفضة التعبير والتى قد تتسبب فى إنخفاض النشاط الإنزيمى enzyme kinetics أثناء الهضم فى وجود تركيز قليل من ماده التفاعل (البروتين). كذلك تؤدى إلى إنخفاض حساسيه الـ mass spectrophotometer تحت هذه الظروف. وجدير بالذكر إن صعوبه تحديد وإظهار البروتينات ذات التركيز المنخفض على الجيل ثنائى الإتجاه 2D gel يمكن إرجاعها إلى وجود بروتينات أخرى ذات تركيز عالى تتداخل مع إظهار تلك البروتينات shadowing of low abundant proteins by high abundant proteins. تبعا لذلك تم تطوير طرق أخرى حديثه للتغلب على بعض هذه المشاكل. البروتينات عاليه التركيز تحجب البروتينات ذات التركيز الأقل على الجيل حيث أن الأول ينتج مناطق صبغ اكبر large spots. ويبدوا أن زياده حجم الجيل قد تقدم حلا لهذه المشكله. بالاضافه إلى إستخدام شرائط IPG تحتوى على مدى ضيق من الـ *pI* (zoom 2D gel) تم تصميمها للتغلب على حجب البروتينات ذات التركيز العالى للبروتينات ذات التركيز الأقل protein shadowing. فى هذا التكنيك يتم إستخدام مجموعه شرائط IPG تحتوى على مدى قصير ومختلف من الـ *pI* مع كل عينه بروتين. فمثلا، بدلا من إستخدام شريط واحد IPG لتغطيه الـ *pI* من ٣-١٠ طوله ١٨ سم فإنه يتم إستخدام من ٤-٥ شرائط كل منها ١٨ سم كل منها يغطى حوالى ١,٥ من مدى الـ *pI*. ينتج من ذلك عديد من الـ 2D gel الكبيره كل منها يغطى مدى معين من الـ *pI* والتى يقل فيها تأثير البروتينات ذات التركيز العالى على إظهار البروتينات قليله التركيز.

٨. ٣. ١. تحليل الكتلة الضوئي Mass Spectrometry.

التطورات والتقدم الحديث فى الـ mass spectrometry مكنت من الإظهار السريع وتحديد تتابع الأحماض الأمينيه الجزئى لكميات قليله من البروتينات. هذه الطرق تعتمد على هضم البروتينات المفصوله من الـ 2D gel إما فى الجيل أو بعد

نقل هذه البروتينات إلى أغشيه. الهضم فى الجيل in-gel digestion وهو الشائع وذلك لسهولته. هذه الطريقه تعتمد على قطع البروتين من الجيل يدويا أو أوتوماتيكيا وهضمها بالتربسين ثم تحديد تتابعها من الأحماض الأمينيه جزئيا. ويتم إستخدام نموذج البتيدات الناتجه من الهضم أو التتابع الجزئى من الأحماض الأمينيه لتحديد الجين الذى ينتج هذا البروتين شكل (٨-٦). التطورات الحديثه فى مجال البروتيومكس تشمل تطبيقات عديده فى مجالات مختلفه من أهمها مجال دراسة الأساس الجزيئى للأمراض وإكتشاف أدويه جديده لعلاج تلك الحالات.

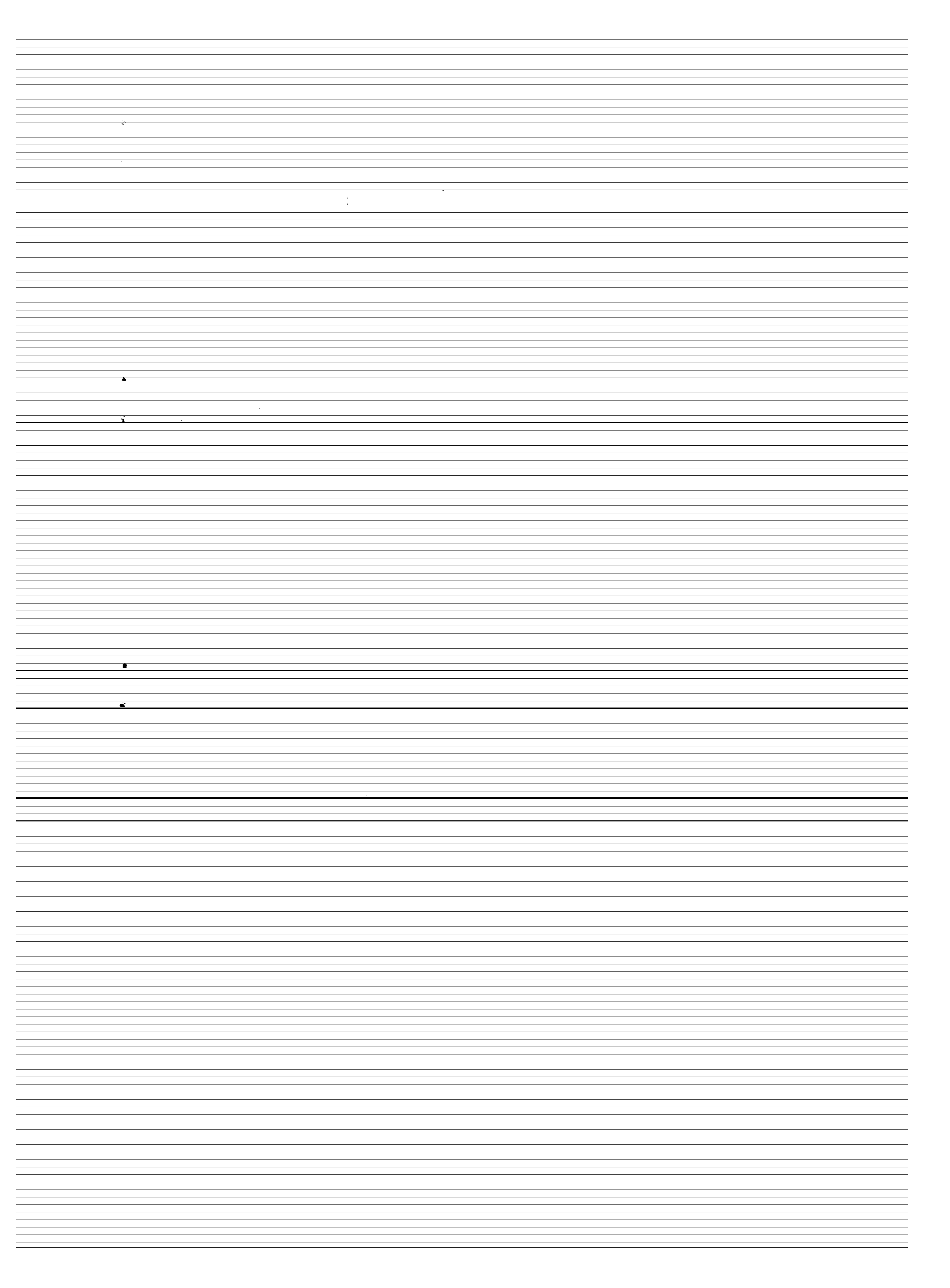


شكل (٨-٦): شكل تخطيطى يوضح إستخدام طريقه تحليل الكتله للتعرف على البروتينات وتحديد تتابع الأحماض الأمينيه لمكوناتها البتيديه.

تستخدم هذه الطريقه لتحديد كتله البروتينات والبتيدات الناتجه بعد تحليلها الإنزيمى بدقه وإستخدام هذه المعلومات للبحث فى قواعد البيانات عن الجينات الخاصه بهذه البروتينات. حيث يتم فصل البروتين من الجيل ثنائى الإتجاه وهضمه بالتربسين ثم

قياس كتله الببتيدات الناتجه من الهضم بواسطه جهاز قياس الكتله. يتم بعد ذلك البحث فى قواعد البيانات عن جينات تنتج بروتينات لها نفس نموذج الهضم بالتربسين.

كذلك يستخدم تحليل الكتله أيضا لتحديد تتابع الأحماض الأمينيه للببتيدات الناتجه من الهضم بالتربسين. فى هذه الحاله فإن البروتينات المفصوله بالـ chromatography كبروتينات مفرده يتم هضمها بالتربسين ثم يتم تحديد كتله الببتيدات الناتجه بجهاز تحديد الكتله بنفس الطريقه كما فى شكل (٨-٦أ). لتحديد التتابع الدقيق للأحماض الأمينيه فإن كل ببتيده ناتجه يتم تحديد تتابعها من الأحمض الأمينيه بتكسير روابطها الببتيديه. هذا ينتج مجموعه من الببتيدات المختلفه فى الطول بحامض أمينى واحد. يتم نقل هذا المخلوط من الببتيدات إلى جهاز تحليل كتله اخر متصل لتحديد كتلتها. الإختلاف فى الكتله بين الببتيدات المختلفه فى حامض امينى واحد تستخدم لتحديد الحامض الأمينى المفقود. بتكرار هذه العمليه يمكن تحديد تتابع الأحماض الأمينيه لجزء من البروتين الكلى (شكل ٨-٦ب).



٩. المراجع References

- أبويوسف، أميرة يوسف و أحمد يوسف المتينى – التقنية الحيوية الجزيئية وتحسين النباتات – منشأة المعارف – الإسكندرية – ٢٠٠٢.
- المتينى، أحمد يوسف – تنظيم الفعل الجينى فى الكائنات الراقية – منشأة المعارف – الإسكندرية – ١٩٩٧.
- المتينى، أحمد يوسف – مدخل الوراثة الجزيئية – منشأة المعارف – الإسكندرية – ١٩٩٤.

---

1. Adams, M. D., M. B. Soares, A. R. Kerlavage, C. Fields, and J. C. Venter, 1993. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nature Genetics, 4: 373 – 380.
2. Antonopoulos, J. E. *Genomics*. Xlibris Corporation, 2000.
3. Barnes, M., and I. C. Gray (Eds.). *Bioinformatics for Geneticists*. John Wiley & Sons, 2003.
4. Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D.L. Wheeler (2005). GenBank. Nucleic Acids Res., 33: 231-244.
5. Bergeron, B. P. *Bioinformatics Computing*. Prentice Hall. 2002.
6. Bishop, M. J., and C. J. Rawlings (Eds.). *DNA and Protein Sequence Analysis: A Practical Approach* (Practical Approach Series, No 171), IRL Press. 1996.

7. Bishop, M. J. (Ed.). *Genetic Databases.* Academic Press, San Diego, 1999.

8. Bornholdt, S., and H. G. Schuster (Eds.). *Handbook of Graphs and Networks : From the Genome to the Internet.* Vch. Verlagsgesellschaft Mbh., 2003.

9. Brown, S. M. *Bioinformatics: A Biologist's Guide to Biocomputing and the Internet.* Eaton Pub.Co., 2000.

10. Campbell, A. C., and L. J. Heyer. *Discovering Genomics, Proteomics, and Bioinformatics.* Benjamin Cummings Pub., 2003.

11. Causton, H. C., J. Quackenbush, and A. Brazma. *Microarray Gene Expression Data Analysis: A Beginner's Guid.,* Blackwell Publishers, 2003.

12. Collado-Vides, J., and R. Hofestadt (Eds.). *Gene Regulation and Metabolism: Post-Genomic Computational Approaches* (Computational Molecular Biology), MIT Press, 2002.

13. Davidson, Erick H. *Genomic Regulatory Systems: Development and Evolution.* Academic Press, San Diego, 2001.

14. Durrett, R., and R. Durrett. *Probability Models for DNA Sequence Evolution.* Springer Verlag, 2002.

15. Eisen, J. A., 1998, Phylogenomics: improving functional prediction for uncharacterized genes by evolutionary analysis. Genome. Res., 8: 163-167.

16. Fall, C., E. Marland, J. Wagner, and J. Tyson (Eds.). *Computational Cell Biology.* Springer Verlag, 2002.

17. Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates, 2003.

18. Fitch, W.M., and Margoliash, E., 1968, The construction of phylogenetic trees. II. How well do they reflect past history? Brookhaven Symp. Biol., 21(1):217-242.

19. Fitch, W.M., and Margoliash, E., 1967, Construction of phylogenetic trees. Science, 155(760):279-284.

20. Fogel, B., and D. W. Corne (Eds.) *Evolutionary Computation in Bioinformatics*, Morgan Kaufmann. 2002.

21. Gingerich, P.D., Haq, Mu., Zalmout, I.S., Khan, I.H., and Malkani, M.S., 2001, Origin of whales from early artiodactyls: hands and feet of Eocene Protocetidae from Pakistan. Science, 293(5538): 2239-2242.

22. Griffiths, A. J..F.; Gelbart, W. M.; Miller, J. H.; Lewontin, R. C. *Modern Genetic Analysis*: W. H. Freeman & Co., 1999.

23. Gribskov, M., and J. Devereox (Eds.). *Sequence Analysis Primer*. Oxford University Press, 1992.

24. Grotewold, E. (Edt). *Plant Functional Genomics* (Methods in Molecular Biology, Vol. 236) (Methods in Molecular Biology (Clifton, N.J.), V. 236.), 2003.

25. Hall, B. G. *Phylogenetics Trees Made Easy: A How-To Manual for Molecular Biologists.* Sinauer Associates, 2001.

26. Hartemink, A. J., D. K. Giford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic

regulatory networks," in Paciuc Symposium on Biocomputing, vol. 6, 2001.

27. Hawley, R. S., and M. Y. Walker. *Advanced Genetic Analysis: Finding Meaning in the Genome.* Blackwell Publishers, 2003.

28. Hedrick, P. W. *Genetics of Populations*. Jones & Bartlett Pub., 2000.

29. Heiskanen, M., Hellsten, E., Kallioniemi, O.P., Makela, T.P., Alitalo, K., Peltonen, L., and Palotie, A., 1995. Visual mapping by fiber-FISH. Genomics. 1;30(1):31-36.

30. Hillis, D. M. In *Homology: the hierarchical bases of comparative biology* (B. K. Hall, Ed.), pp. 339-368, Academic Press, San Diego 1994.

31. Hunt, S., and F. Livesey (Eds.). *Functional Genomics: A Practical Approach* (The Practical Approach Series, 235), Oxford Univ. Press, 2000.

32. James, P. (Ed.). *Proteome Research: Mass Spectrometry (Principles and Practice)*, Springer Verlag, 2001.

33. Kimura, M., and N. Takahata (Eds.). *Population Genetics, Molecular Evolution, and the Neutral Theory*: Selected Papers. University of Chicago Press. 1964.

34. Koonin, E. V., and Galperin, M. Y. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, 2002.

35. Koski, T., and T. Koskinen. *Hidden Markov Models for Bioinformatics*, Kluwer Academic Publishers. 2001.

36. Klug, W., and M. R. Cummings. *Concepts of Genetics*., Fifth Edition, Printic Hall Inc., 1999.

37. Lesk, A. M. *Introduction to Bioinformatics*. Oxford University Press. 2002.

38. Letovesky, S. (Ed.). *Bioinformatics: Databases and Systems*. Kluwer Academic Publishers. 1999.

39. Levine, M. and Eric H. Davidson, 2005. Gene regulatory networks for development. Proc. Natl. Acad. Sci., USA, 102: 4936-4942.

40. Liebler, D. C. (Ed.). *Introduction to Proteomics: Tools for the New Biology*. Humana Press, 2001.

41. Nei, M. ,1996, Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet., 30:371-403.

42. Nei, M., and S. Kumar. *Molecular Evolution and Phylogenetics*. Oxford Univ Press, 2000.

43. Ou, C.Y., Ciesielski, C.A., Myers, G., Bandea, C.I., Luo, C.C., Korber, B.T., Mullins, J.I., Schochetman, G., Berkelman, R.L., Economou, A.N.,1992, Molecular epidemiology of HIV transmission in a dental practice. Science, 256(5060):1165-1171.

44. Palzkill, T. *Proteomics*. Kluwer Academic Publishers, 2002.

45. Percus, J. K. (Ed.). *Mathematics of Genome Analysis*. Cambridge Univ. Press. 2002.

46. Rashidi, H. (Ed.). *Bioinformatics Basics: Applications in Biological Science and Medicine*. CRC Press. 1999.

47. Rastrigin, L. *This Chancy, Chancy, Chancy World*. Mir Publishers, 1973.

48. Ross, S. *A First Course in Probability* (6th Edition), Prentice Hall, 2002.

49. Rothstein, M. A. (Ed.). *Pharmacogenomics: Social, Ethical, and Clinical Dimensions*. Wiley-Liss, 2003.

50. Saccone, C., and G. Pesole. *Handbook of Comparative Genomics: Principles and Methodology*. Wiley-Liss, 2003.

51. Salemi,M., and Anne-Mieke Vandamme (Eds.). *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*. Cambridge University Press, 2003.

52. Sankoff, D., and J. H. Nadeau. *Comparative Genomics - Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*. Kluwer Academic Pub., 2000.

53. Sensen, C. W. (Ed.). *Essentials of Genomics and Bioinformatics*. John Wiley & Sons. 2002.

54. Sorensen, D. and D. Gianola. *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer Verlag, 2002.

55. Suhat, S. (Ed.). *Genomics and Proteomics: Functional and Computational Aspects*. Plenum Pub. Corp., 2000.

56. Till, B., Steven H. Reynolds, Clifford Weil, Nathan Springer, Chris Burtner, Kim Young, Elisabeth Bowers, E., C. A. Codomo, L. C. Enns, A. R. Odden, E. A. Greene, L. Comai and S. Henikoff , 2004. Discovery of induced point mutations in maize genes by TILLING. BMC Plant Biology, 4:12-17.

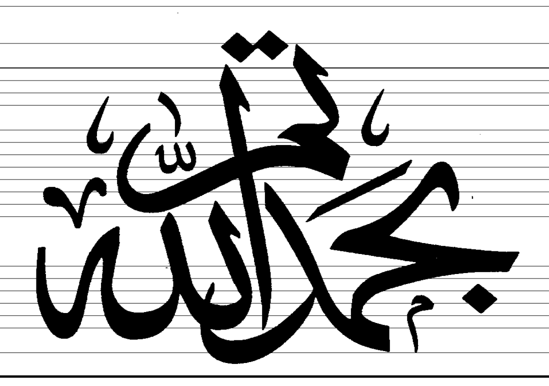
57. Thewissen, J.G., Williams, E.M., Roe, L.J., and Hussain, S.T., 2001, Skeletons of terrestrial cetaceans and the relationship of whales to artiodactyls. Nature, 413(6853):277-281.

58. Venter, J. Craig *et al.*, 2001, The Sequence of the Human Genome. *Science*, l291 (5507):1304-1351.

59. Wang, J. T. L., Shapiro, B. A., Shasha, D. E. (Eds.). *Pattern Discovery in Biomolecular Data: Tools, Techniques, and Applications*. Oxford Univ Press. 1999.

60. Waterman, M. S. (Ed.). *Mathematical Methods for DNA Sequences*. CRC Press, 1989.

61. Wren, B., and N. Dorrell (Eds.). *Functional Microbial Genomics*. (Volume 33). Academic Press, 2003.



مكتبة بستان المعرفة
لطباعة ونشر وتوزيع الكتب
كفر الدوار - الحدائق - بجوار نقابة التطبيقيين
☎: ٠٤٥/٢٢٢٤٢٢٨ & ٠١٢١١٥١٢٣٧